پرس و جو روی داده های رمز شده

Masood Niazi Torshiz
www.mniazi.ir

مقدمه

- روش‌های کنترل دسترسی برای محافظت از داده ها کافی نیستند
  - سرقت رسانه محتوی داده
  - عدم اعتماد به اعمال کننده خط مشی های کنترل دسترسی
  - امکان دور زدن مکانیزم های کنترل دسترسی توسط مهاجمین

- مطرح شدن ایده Database as A Service و سیستم های کارگزار غیرقابل اعتماد

مدل پایگاه داه به عنوان خدمت

- پایگاه داده به عنوان خدمت (Database as A Service) به عنوان رویکردی جدید در برون‌سپاری پایگاه داده‌ها
- در دسترس بودن داده ها توسط کارگزار تضمین می شود.
- کلیه اعمال مدیریت داده را کارگزار فراهم می کند.
- کارگزار از نظر نگهداری داده‌ها و عدم ارسال عمدی پاسخ اشتباه مورد اعتماد است.
- کارگزار در مورد محرمانگی داده‌ها مورد اعتماد نیست.
  - کارگزار درستکار ولی کنجکاو است (Honest but curious).

مدل پایگاه داده به عنوان خدمت

- چالش اصلی در این مدل تأمین امنیت داده‌های برون‌سپاری شده است.
- راه حل اولیه رمزنگاری داده های برون‌سپاری شده است.
- برای حفظ محرمانگی مالک داده، داده خود را رمز کرده و آن را در پایگاه داده رمز شده در سمت کارگزار ذخیره می کند.
- ریزدانگی رمزنگاری به خط مشی های محیط برای سطح دسترسی، امنیت و کارایی بستگی دارد.
  - بیشتر فعالیت ها ریزدانگی را در سطح چندتایی تعریف کرده اند.

عناصر مدل DAS

1. مالک داده‌ها: فرد یا سازمان است که داده‌ها را ایجاد و آن را برون‌سپاری می‌کند.

2. کاربر: پرس‌وجوها را به سیستم ارائه می‌کند.

3. کارخواه: پرس‌وجوهای کاربر را به پرس‌وجوهای قابل اجرا روی داده‌های رمزشده تبدیل می‌کند.

4. کارگزار: محل ذخیره‌ی داده‌های رمز شده است و پرس‌وجوهای ارسالی از سمت کارخواه را روی داده‌های رمزشده اجرا کرده و نتیجه را به کارخواه ارائه می‌دهد.

سناریوی پرس و جو در مدل DAS

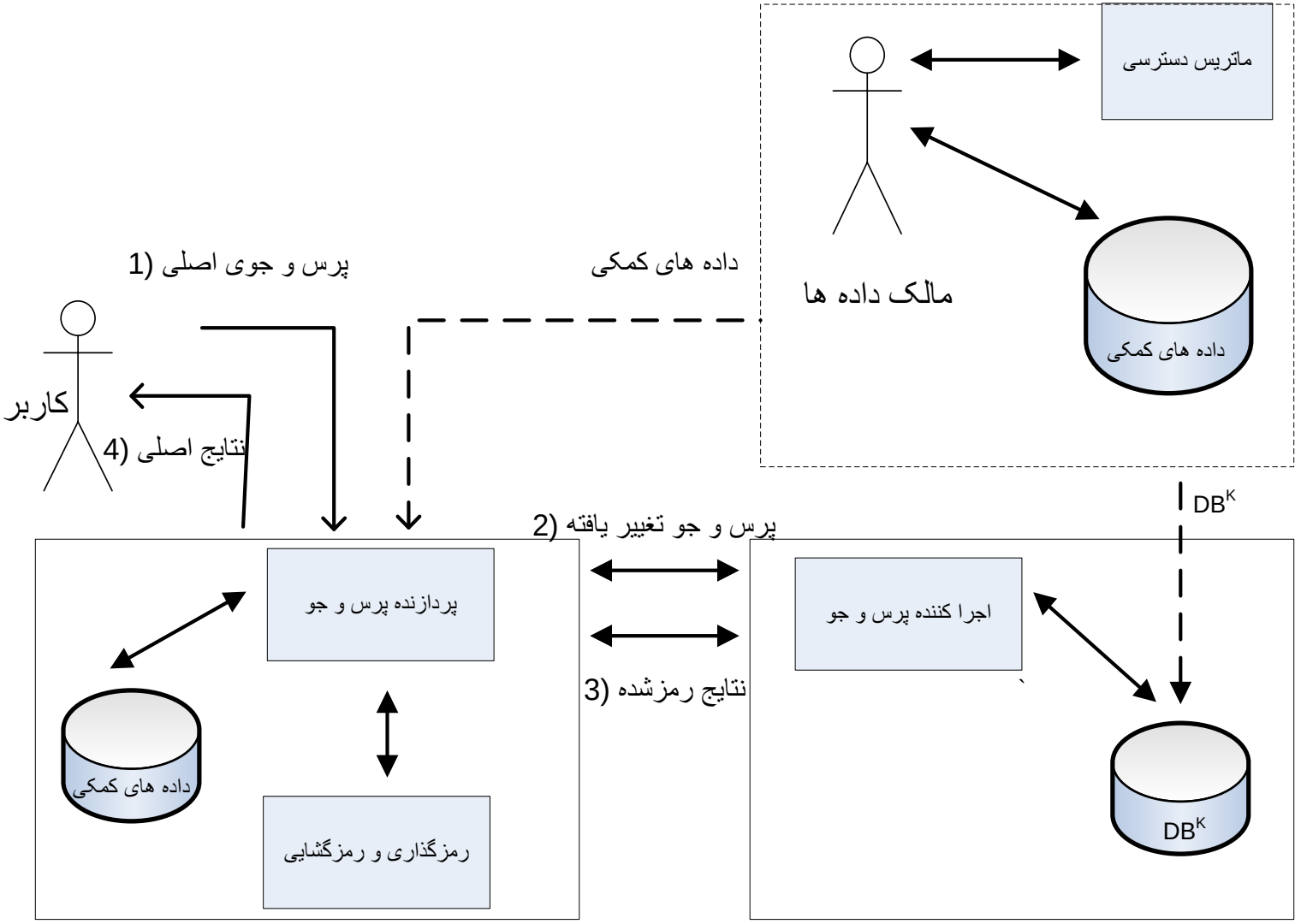
1. کاربر پرس و جوی Q را با توجه به شمای پایگاه داده ی رمز نشده B از طریق کارخواه وارد می کند.
   - برون سپاری داده می تواند از دید کاربر شفاف باشد.

2. کارخواه پرس و جوی کاربر را به دو بخش Qs و Qc تقسیم می کند. Qsپرس و جوی اعمال شده بر روی داده های رمز شده در سمت کارگزار و Qc پرس و جوی اعمال شده در سمت کارفرما بر روی داده های برگشتی از کارگزار به کارخواه است.
   - کارخواه ساختار پایگاه داده عادی و رمز شده را می داند

3. کارگزار پرس و جوی Qs را روی داده رمز شده اجرا و نتایج (مجموعه ای از چندتایی های رمز شده) را به کارخواه بر میگرداند.

4. کارخواه نتایج را رمزگشایی کرده و چندتایی های اضافی را با اعمال Qc به نتایج اولیه حذف می کند. نتایج نهایی به کاربر ارائه می شود.

سناریوی پرس و جو در مدل DAS



ملاحظات رمزنگاری در برون سپاری داده

- روش هایی که بتوانند به طور مستقیم با داده های رمز شده کار کند باید ملاحظات زیر را درنظر بگیرند:
  - میزان اعتماد به کارگزار
    - در مدل DAS امکان رمزگشایی توسط کارگزار نامطمئن وجود ندارد.
  - کارایی روش اجرای پرس و جو
    - رمزگشایی کل داده های قبل از اجرای پرس و جو کارا نیست.
  - تمرکز اجرای اعمال در سمت کارگزار
  - سربار قابل قبول برای ذخیره سازی و ارتباطات بین کارفرما و کارگزار
  - ریزدانگی رمزنگاری
    - اگر رمزنگاری بصورت درشت‌دانه باشد امکان بهینه‌سازی پرس‌وجو کم می شود
    - رمزنگاری به صورت ریز دانه نیز کارایی را کمتر و در شرایطی به ممکن است به مهاجم اجازه استنتاج از داده‌ها را بدهد.
  - کنترل دسترسی در سیستم های چند کاربره

ملاحظات رمزنگاری در برون سپاری داده (۲)

- مقاومت در برابر حملات

- حمله متن رمزشده معلوم: به طور کلی فرض می‌شود که مهاجم به داده رمزشده دسترسی دارد. هدف در این حمله شکستن متن رمزشده خاص یا پیدا کردن کلید است.
- حمله متن اصلی معلوم: مهاجم به تعدادی متن اصلی و معادل رمزشده آن‌ها دسترسی دارد که از آن برای به دست آوردن بقیه‌ی متون رمزشده یا پی بردن به کلید رمز استفاده می‌کند.
- حمله متن اصلی انتخابی: مهاجم می‌تواند معادل رمزشده متن اصلی دلخواه خود را به دست بیاورد. این حمله، نوع قویتری نسبت به حمله‌ی متن اصلی معلوم است.
- حمله متن رمز شده انتخابی: مهاجم می‌تواند رمزگشایی شده معادل متن رمزشده دلخواه را بدست آورد.
- حملات تحلیل فرکانسی: ممکن است مهاجم (server) اطلاعات اولیه‌ای راجع به دامنه مقادیر و فرکانس رخداد داده‌های رمزنشده داشته باشد و از آن برای نفوذ به پایگاه داده استفاده کند.
- حملات مبتنی بر اندازه: ممکن است مهاجم اطلاعاتی راجع‌به ارتباط طول متن اصلی و متن رمزشده داشته باشد. بنابراین اگر مهاجم مجموعه‌ای از داده‌های اصلی و متن رمز شده معادل را داشته باشد می‌تواند به پایگاه داده حمله کند.

ملاحظات رمزنگاری در برون سپاری داده (۳)

- پشتیبانی از انواع پرس و جو
  - پرس و جو روی داده های عددی
    - پرس و جو با شرط تساوی
    - پرس و جوی بازه ای
  - پرس و جو روی داده های رشته ای
    - پرس و جو با شرط تساوی
    - پرس و جو های تطبیق الگویی
  - پرس و جوهای شامل توابع تجمعی

روش های جستجو روی داده های رمز شده

- جستجوی مستقیم روی داده های رمز شده
- جستجوی مبتنی بر شاخص
- روش های مبتنی بر حفظ ترتیب
- روش های مبتنی بر توابع همریخت اختفایی

جستجوی مستقیم روی داده های رمز شده

- داده به گونه ای رمز می شود که جستجو بتواند دقیقاً روی همان داده رمز شده به صورت مستقیم صورت گیرد.

- سانگ روشی را بر اساس این ایده برای جستجو روی داده های رشته ای ارائه داده است.

روش Song – معرفی

- جستجوی کلمات روی اسناد رمز شده (تمرکز بر DB نیست)
- کاربرد مفهوم دریچه
- کارگزار می‌تواند با گرفتن اطلاعات کوچکی در مورد هر کلمه (دریچه)، جستجو را بدون اطلاع از کلمات دیگر متن انجام دهد.
- توابع مبتنی بر دریچه توابعی هستند که محاسبه‌ی معکوس آنها بدون داشتن اطلاعات خاصی به نام دریچه مشکل است.
- در رمزنگاری مبتنی بر دریچه، رمزگشایی با داشتن دریچه امکان‌پذیر است.
- در این روش‌ها، به همراه هر کلمه‌ای که کارخواه جستجوی آنرا تقاضا کرده است، دریچه‌ی آن نیز ارسال می‌شود. بدین شکل کارگزار فقط می‌تواند کلمه درخواست شده را رمزگشایی کند.

روش Song – رمزگذاری

1. متن اصلی به تعدادی کلمه w با طول یکسان (n بیت) تقسیم می شود.
2. اسناد اصلی پس از رمزشدن به روش شرح داده شده، به سمت کارگزار ارسال و در آنجا ذخیره می‌شوند.
3. کارگزار با دریافت دریچه‌ای از طرف کارخواه می تواند کلمه‌ی مورد نظر کاربر را جستجو کند.

- پارامترهای رمزنگاری
  - S: مولد اعداد شبه تصادفی
  - F و f: توابع شبه تصادفی
  - K’: کلید تابع f (برای تمام کلمات متن ثابت است)
  - دریچه هر کلمه: کلید تابع F: $f_{k'}(\text{first n-m bits of } E_{k''}(w_i))$

روش Song – رمزگذاری(۲)

- رمزگذاری در دو سطح انجام می شود.
  - سطح اول:
    - هر کلمه با یکی از الگوریتم‌های رمزنگاری متقارن و کلید ("k") رمز می شود.
    - کارگزار در هنگام اجرای پرس‌وجو از کلمه‌ی درخواست شده کارخواه مطلع نمی شود.
  - سطح دوم:
    - مولد شبه تصادفی S، دنباله‌ای از اعداد شبه تصادفی s_i با طول n-m بیت به تعداد کلمات متن اصلی ایجاد می‌کند.
    - اعداد شبه تصادفی تولید شده s_i با استفاده از تابع F درهم‌سازی شده و خروجی m بیتی تولید می شود.
    - (رمزشده‌ی لایه‌ی اول هر کلمه ($E_{k''}(w_i)$) با (s_i و حاصل درهم‌سازی شده در مرحله قبل)، XOR می‌شود.
    - نتیجه‌ی لایه‌ی دوم رمزنگاریِ کلمه‌ی w_i به عنوان iامین کلمه‌ی متن رمزشده (c_i) در سند رمزشده قرار می‌گیرد.

روش Song – رمزگذاری(۳)

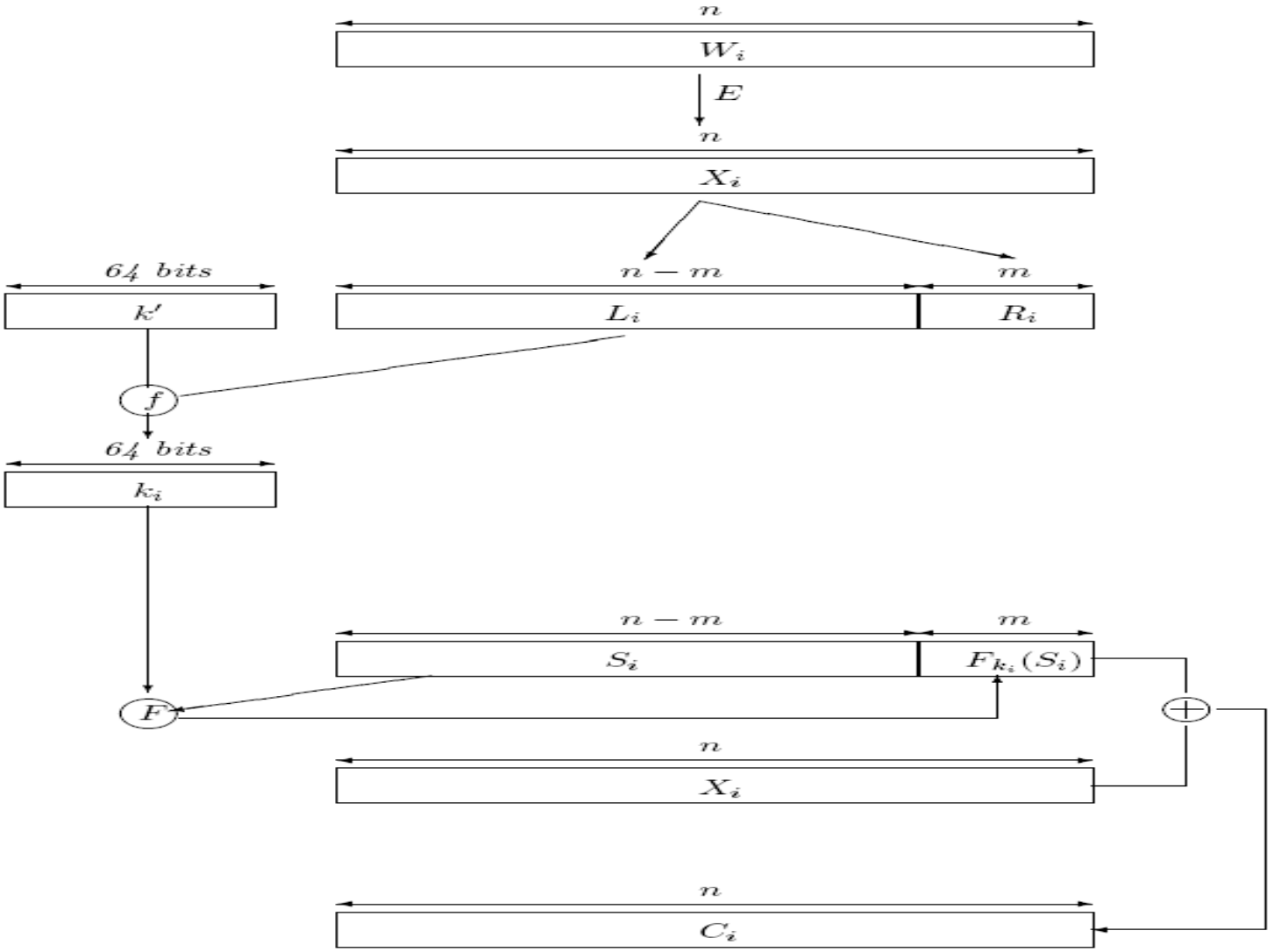


Figure 1: Encryption schema

روش Song – اعمال پرس و جو

1. کارخواه برای جستجوی یک کلمه (w) در اسناد رمزشده، معادل رمز شده‌ی لایه‌ی اول کلمه ($E_{k''}(w)$) به همراه دریچه‌ی آن ($f_{k'}(w)$) را به کارگزار ارسال می‌کند.
2. کارگزار با دریافت ($E_{k''}(w)$) کلمات تمام اسناد را با آن XOR می‌کند.
3. اگر کلمه‌ی Pام سندی با کلمه‌ی درخواست شده برابر باشد، حاصل XOR (T_p) باید ساختاری به شکل $<S_p, F_{f_{k'}(w)}(s_p)>$ داشته باشد. برای بررسی وجود ساختار فوق برای کلمه‌ی pام متن رمزشده، حاصل تابع F روی n-m بیت پرارزش T_p به دست آورده می‌شود.
4. اگر مقدار به دست آمده با m بیت باقی‌مانده‌ی T_p برابر باشد، ساختار برقرار بوده و کلمه‌ی Pام متن رمزشده به همراه سندی که به آن متعلق است در مجموعه‌ی جواب ارسالی به کارخواه قرار می‌گیرد.

- تابع F یک تابع درهم‌ساز دارای برخورد است. بنابراین امکان وجود اشتباه مثبت در نتایج ارسالی به کارخواه وجود دارد.
  - در سمت کارخواه پس از رمزگشایی سند، مقدار اصلی کلمه‌ی پیدا شده با کلمه‌ی درخواست شده‌ی کاربر مقایسه می‌شود تا نتایج درستی به کاربر برگردانده شود.

روش Song – رمزگشایی

- برای رمزگشایی کلمه‌ی *i*اُم سند رمزشده (c_i)، ابتدا $n-m$ بیت پرارزش c_i با s_i، XOR می‌شود و $n-m$ بیت پر ارزش $E_{k''}(w_i)$ به دست می آید.
- از مقدار فوق برای ساختن دریچه‌ی w_i استفاده می‌شود.
- اعمال دریچه بدست آمده و n-m بیت پر ارزش s_i به تابع F، m بیت نتیجه دارد که با XOR کردن با m بیت کم ارزش c_i، m بیت کم ارزش $E_{k''}(w_i)$ حاصل می شود. بدین ترتیب تمام بیت‌های $E_{k''}(w_i)$ به دست می‌آیند.
- $E_{k''}(w_i)$ رمزگشایی شده تا مقدار اصلی w_i حاصل شود.

ویژگی های روش Song

- کارگزار نمی تواند در مورد متن اصلی تنها با استفاده از متن رمز شده اطلاعاتی بدست آورد.
- سربار ذخیره سازی و ارتباطاتی آن کم است.
- نتیجه حاوی مکان هایی از سند است که W در آن ظاهر شده است و ممکن است دارای اشتباهات مثبت باشد.
- اشتباهات مثبت با مقدار m مرتبط است. هر جواب اشتباه با احتمال $1/2^m$ رخ می دهد. بنابراین برای سندی با طول l کلمه انتظار $l/2^m$ جواب اشتباه وجود دارد.
- متن باید تقسیم به کلماتی با طول مساوی شود که با توجه به ساختار زبان، روش مناسبی نیست.

ویژگی های روش Song (۲)

- امکان جستجو با هر طول دلخواه وجود ندارد. فقط می‌توان کلمات با طول n یا ضریبی از n بیت را جستجو کرد.
- گروه محدودی از الگوها قابل جستجو است.
  - الگوهایی به شکل *"ab[a-z]"* با تبدیل به *aba, abb, abc, ..., abz* قابل جستجو هستند؛
  - جستجوی الگوهایی به شکل *"ab*"* مشکل است. زیرا تعداد رشته‌های تولیدی بسیار زیاد خواهند شد.
- در الگوریتم سانگ، برای یافتن هر کلمه باید کل محتویات تمام اسناد جستجو شود. زمان جستجو نسبت به طول متن خطی است. بنابراین در مقیاس بزرگ (مانند پایگاه داده) کارا نیست.
  - یک روش افزایش سرعت بکارگیری شاخص های از پیش تعریف شده است.

جمع بندی - جستجوی مستقیم روی داده رمز شده

- فضای ذخیره سازی نسبت به ذخیره سازی داده اصلی تفاوت زیادی ندارد.
- معمولاً احتمال اجرای حملات فرکانسی نسبت به روش مبتنی بر شاخص کمتر است.
  - در برخی روش ها کلمه ای که چندین بار بکار رفته، هر بار به شکل جدیدی رمز می شود (بسته به جایگاه کلمه در متن)
- بیشتر اجرای پرس و جو در سمت کارگزار است. تنها رمزنگاری کلمه مورد پرس و جو در سمت کارخواه انجام می شود.
- نسبت به روش های مبتنی بر شاخص دارای جواب اشتباه کمتری است.
- پیچیدگی محاسباتی بالایی دارند و زمان جستجو در آن ها خطی است. بنابراین در مقیاس بزرگ قابل استفاده نیست.
- این روش ها بیشتر برای ابزارهایی با جستجوی در مقیاس کوچک (مانند تلفن همراه) مناسب است.
- پرس و جوهای با شرایط تطبیق دقیق را پاسخ می دهند.
- پرس وجوی شامل شرایط بازه ای و جستجوی الگوها بر روی داده های رشته ای در این روش سخت است.
- پرس و جو شامل توابع تجمعی امکان پذیر نیست.

جستجوی مبتنی بر شاخص

- اطلاعاتی با نام شاخص همراه با داده رمز شده در سمت کارگزار ذخیره می شود.
- جستجوی داده با استفاده از شاخص ذخیره شده انجام می گیرد.
- برای هر عنصر که بخواهد جستجو بر مبنای آن انجام شود، باید یک شاخص تعریف کرد.
- شاخص نباید اطلاعاتی در مورد داده اصلی را فاش نماید.
- روش های مختلف تولید شاخص باید از یک سو کارایی پرس و جو و از سوی دیگر عدم سوءاستفاده کارگزار از مقدار شاخص در استنتاج داده اصلی را در نظر بگیرند.

جستجوی مبتنی بر شاخص

- R_i $(A_{i1}, A_{i2}, ..., A_{in})$ در پایگاه داده رمز نشده به R^k_i (Counter, Etuple, I_1, I_2, . . ., I_n) در پایگاه داده رمز شده نگاشت می شود. Counter کلید اصلی جدول رمز شده، Etuple رمز شده چندتایی معادل در پایگاه داده رمزنشده، I_1 تا I_n شاخص های متناظر با A_{i1} تا A_{in} هستند.

Employee

Emp-Id	Name	YoB	Dept	Salary
P01	Ann	1980	Production	10
R01	Bob	1975	R&D	15
F01	Bob	1985	Financial	10
P02	Carol	1980	Production	20
F02	Ann	1980	Financial	15
R02	David	1978	R&D	15

Employee^k

Counter	Etuple	I_1	I_2	I_3	I_4	I_5
1	ite6*+8wc	π	α	γ	ε	λ
2	8(nfeua4!=	ϕ	β	δ	θ	λ
3	Q73gnew321*/	ϕ	β	γ	μ	λ
4	-1vs9e892s	π	α	γ	ε	ρ
5	e32rfs4+@	π	α	γ	μ	λ
6	r43arg*5[)	ϕ	β	δ	θ	λ

روش های اصلی مبتنی بر شاخص

- Bucket Based Index
- Hash Based Index
- B+ Tree Index

شاخص مبتنی بر Bucket- معرفی

- هاسیگموس و همکارانش مبتنی بر افزودن شاخص مبتنی بر باکت، روشی برای پرس و جو روی داده رمز شده ارائه دادند.
- ردیف‌های هر جدول به طور جداگانه و به کمک یکی از الگوریتم‌های رمزنگاری متقارن، رمز می‌شوند.
- به ازای هر مشخصه که جستجو روی آن انجام می‌شود، یک مشخصه به نام شاخص به جدول اضافه می‌گردد.

ID	***Title***	***Author***	***Year***
123	*Handbook of Applied Criptography*	*Alfred J. Menezes*	*2001*

Enc_Tuple	I_{ID}	I_T	I_A	I_Y
4%n+!~kl?7klm||fcapkmvk380($%	Σ	Π	Φ	Ψ

نحوه تعریف شاخص در روش Bucketبندی

1. برای ایجاد شاخص امن، بازه‌ی داده‌های هر صفت به تعدادی زیربازه تقسیم می‌شود.
   - بازه ها نباید اشتراک داشته باشند.
2. ایجاد بازه‌ها می‌تواند با استفاده از انواع روش‌های تقسیم‌بندی صورت گیرد.
   - طول برابر
   - تعداد اعضای برابر
3. به هرکدام از بازه‌ها مقداری تعلق می‌گیرد.
   - می توان از یک تابع تولید اعداد شبه تصادفی استفاده کرد.

- به ازای تمام داده‌هایی که در یک زیربازه قرار می‌گیرند، مقدار تعلق گرفته به زیربازه‌ی آنها در شاخص رمزشده قرار داده می‌شود.

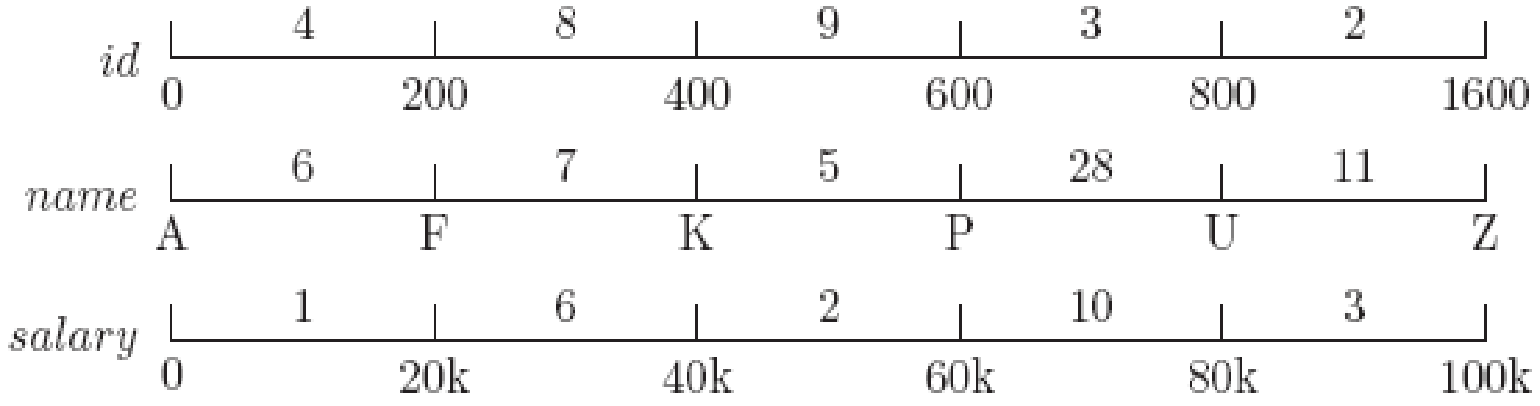
مثال – بازه بندی و اعطای شاخص

id	*name*	*salary*
23	Tom	70000
860	Mary	60000
320	Tony	50000
875	Jerry	5600

Table 1.1: Plain text *salary* table.

etuple	id^S	$name^S$	$salary^S$
010101011...	4	28	10
000101101...	2	5	10
010111010...	8	28	2
110111101...	2	7	1

Table 1.2: Encrypted *salary* table.



پرس‌وجو در شاخص دهی مبتنی بر Bucket

- پرس‌وجو در دو مرحله صورت می‌گیرد:
  1. کارگزار سعی می‌کند تا آنجا که امکان دارد پاسخ درستی برگرداند
  2. کارخواه نتیجه‌ی ارسالی کارگزار را رمزگشایی کرده و آنرا پردازش می‌کند.
- در SQL یک پرس و جو می تواند با درخت‌های متفاوت که از نظر نتیجه یکسان هستند، نمایش داده شود.
- هاسیگموس جداسازی پرس و جو را (بخشی در سمت کارگزار و بخشی در سمت کارخواه) را با استفاده از درخت پرس‌وجو انجام داده است.

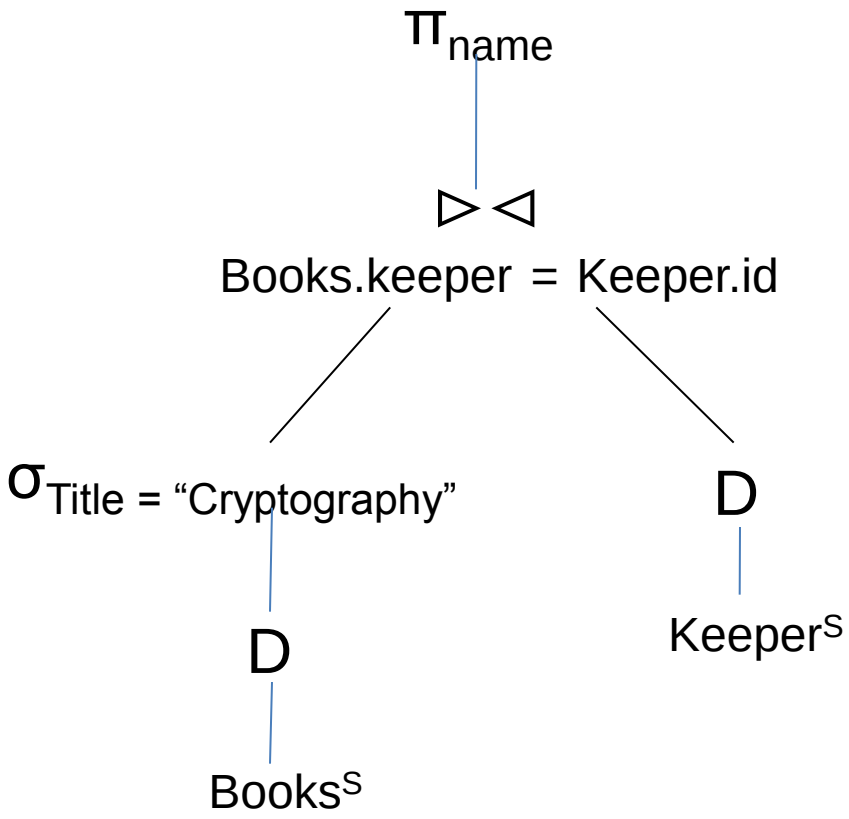
پرس و جو در شاخص دهی مبتنی بر Bucket (۲)

- در روش هاسیگموس، درخت نمایشِ هر پرس‌وجو به دو قسمت تقسیم می‌شود.
  1. بخش زیر عملگر رمزگشایی: کلیه اعمالی که می تواند در سمت کارگزار انجام شود.
  2. بخش بالای عملگر رمزگشایی: اعمالی که باید در سمت کارخواه انجام شوند.
- اَعمال جبر رابطه‌ای وقتی به زیر عمل رمزگشایی برده می‌شوند، باید به اَعمال جدیدی که روی داده‌های رمزشده قابل اجرا هستند، تبدیل شوند.
  - هاسیگموس جبر رابطه‌ای جدیدی را برای اِعمال روی داده‌های رمزشده تعریف کرد.

بهینه سازی اجرای پرس و جو

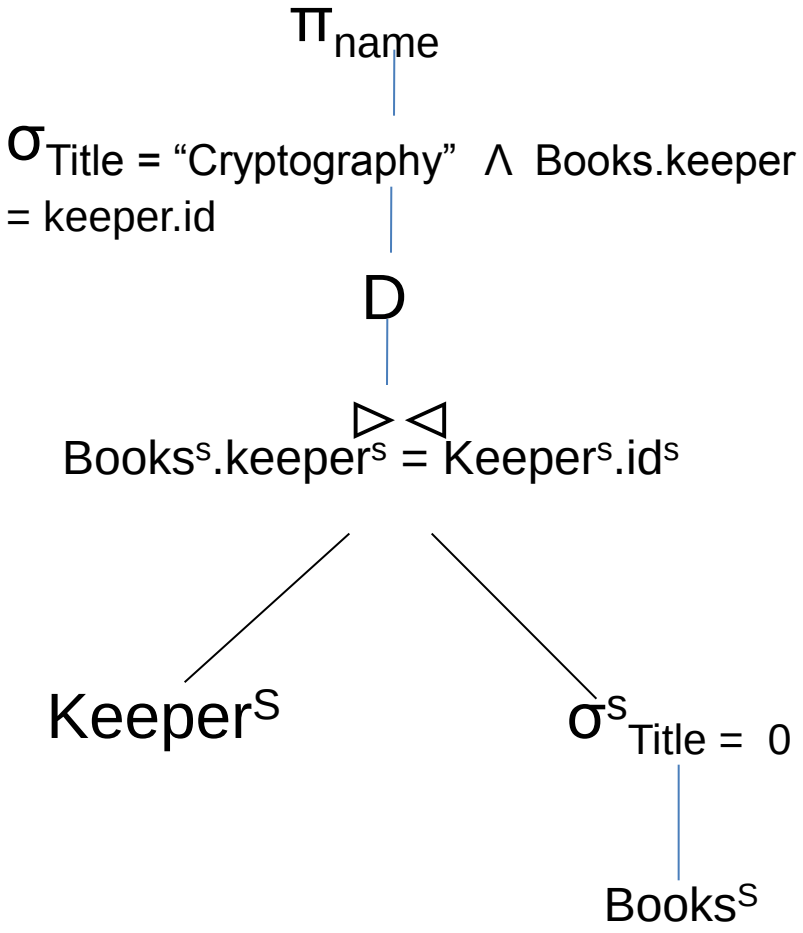
SELECT name ***FROM*** books, keeper

WHERE keeper.ID = books.keeping ***AND*** Title = "*Handbook of applied cryptography*"



- **فقط عملگرهایی که در زیر عملگر رمزگشایی هستند قابل اجرا در سمت کارگزار هستند**
- **در این اجرا هیچ عملگری قابل اجرا به روی کارگزار نیست.**
- **سربار ارتباطي و محاسباتي در سمت کارخواه بالا است.**
- **براي افزايش کارآيي بايد سعي کرد تا حد امکان، عملگرهاي رابطه‌اي را به زير عملگر رمزگشايي منتقل کرد.**

اجرای بهینه پرس و جو



- اَعمال تا حد امکان باید در سمت کارگزار انجام شود.
- رابطه *SELECTION* روی جدول *Books* به زیر عملگر رمزگشایی برده شده است.
- برای بردن این عملگر به زیر عملگر رمزگشایی باید آنرا تبدیل به σ^s که در جبر جدید هاسیگموس برای *SELECTION* روی جداول رمزشده و مشخصه‌های شاخص تعریف شده است، تبدیل کرد.
- این اجرا از پرس‌وجو به دلیل اجرای برخی از عملگرها در سمت کارگزار اجرای کارایی است.

مزایا و معایب روش باکت بندی برای تولید شاخص

- روش های مبتنی بر Bucket برای اجرای شروط تساوی مناسب هستند.
  - $Aij = v \rightarrow Ij = \beta$
- پرس و جوهای بازه ای نیز با کمی تغییر در این روش شاخص دهی قابل اجرا است.
  - $A_{ij} > v \rightarrow (I_j = \beta_1 \text{ or } \beta_2 \text{ or } \ldots \text{or } \beta_k)$
- مثال:
  - شرط Salary>50000 در جدول صفحات قبل باید به شکل زیر تبدیل شود:

    $\text{Salary}^s=2 \text{ OR Salary}^s=3 \text{ OR Salary}^s=10$
- عدم امکان اجرای توابع تجمعی مانند SUM، MIN، MAX، Avg و ...
  - هاسیگموس این روش را تعمیم داد و با استفاده از یک تابع همریخت اختفائی اجرای توابعی مثل جمع و ضرب را فراهم کرد.

مزایا و معایب روش باکت بندی برای تولید شاخص (۲)

- وجود مقدار زیادی از داده های اضافی در نتایج کارگزار (به ویژه اگر اعمالی مانند پیوند در پرس و جو وجود داشته باشد)
  - روش‌هایی برای بهبود و کاهش اشتباه‌های مثبت در روش هاسیگموس ارائه شده است.
- این روش بیشتر برای داده‌های عددی کاربرد دارد و به کارگیری روش برای داده‌های رشته‌ای یا متنی که دامنه‌ی داده‌ها بزرگ است، چندان مناسب نیست.
  - در تعریف بازه برای داده‌های رشته‌ای، ممکن است تعداد زیادی داده در یک بازه قرار گیرد که این خود به معنای ارسال تعداد زیادی داده‌ی اضافی در پاسخ به یک پرس‌وجو و در نتیجه سربار ارتباطی و محاسباتی بالا است.
- اگر بازه ها طوری ساخته شوند که مثلاً در هر بازه یک مقدار قرار بگیرد امکان تحلیل فرکانسی وجود دارد.

شاخص مبتنی بر توابع درهم ساز

- از مفهوم توابع درهم ساز برای ساختن شاخص استفاده می شود.
- برای ساختِ شاخص در این روش، داده‌ها توسط یک تابع درهم‌سازِ دارای تصادم دسته‌بندی می‌شوند.
- مزیت این نوع شاخص دهی نسبت به روش باکت بندی این است که دسته‌بندی روی داده‌های پشت سرهم انجام نمی‌شود که این می‌تواند باعث افزایش امنیت گردد.

شاخص مبتنی بر توابع درهم ساز (۲)

- اکر r_i رابطه ای با شمای $R_i(A_{i1}, A_{i2}, \ldots, A_{in})$ و r^k_i رابطه متناظر رمز شده آن باشد، برای هر صفت A_{ij} در R_i که بخواهیم برای آن شاخص تعریف کنیم، تابع درهم ساز یک طرفه $h: D_{ij} \rightarrow B_{ij}$ تعریف می شود که D_{ij} دامنه A_{ij} و B_{ij} دامنه شاخص I_j متناظر با A_{ij} است.
- for all x, y in D_{ij}; if x = y then h(x)= h(y)
- برد h از دامنه آن کوچکتر است .
  - امکان برخورد وجود دارد.
- برای هر دو مقدار تصادفی متفاوت ولی نزدیک به هم x و y در دامنه D_{ij} $(|x - y| < \varepsilon)$، توزیع احتمالی h(x) – h(y) یکنواخت است.
  - تابع درهم ساز ترتیب خصیصه در دامنه را حفظ نمی کند.

شاخص مبتنی بر توابع درهم ساز – ویژگی ها

- اجرای پرس و جوهای تساوی (مانند روش های شاخص دهی مبتنی بر bucket) ممکن است.
  - هر شرط Aij=v به شرط lj=h(v)، که lj شاخص متناظر با Aij در جدول رمز شده است، تبدیل می شود.
- برخورد در توابع درهم ساز باعث ارسال چندتایی های اضافی به سمت کارخواه می شود.
  - تابع درهم ساز فاقد برخورد مشکل ارسال نتایج زیادی به کارخواه را رفع می کند ولی احتمال تحلیل های فرکانسی را افزایش می دهد.
- مشکل اصلی روش های شاخص دهی مبتنی بر توابع درهم ساز عدم پشتیبانی از پرس و جوهای بازه ای است.
  - دامیانی و همکارانش برای اضافه کردن قابلیت انجام جستجوی بازه‌ای به این روش از شاخص مبتنی بر درخت B+ در کنار این نوع شاخص دهی استفاده کردند.

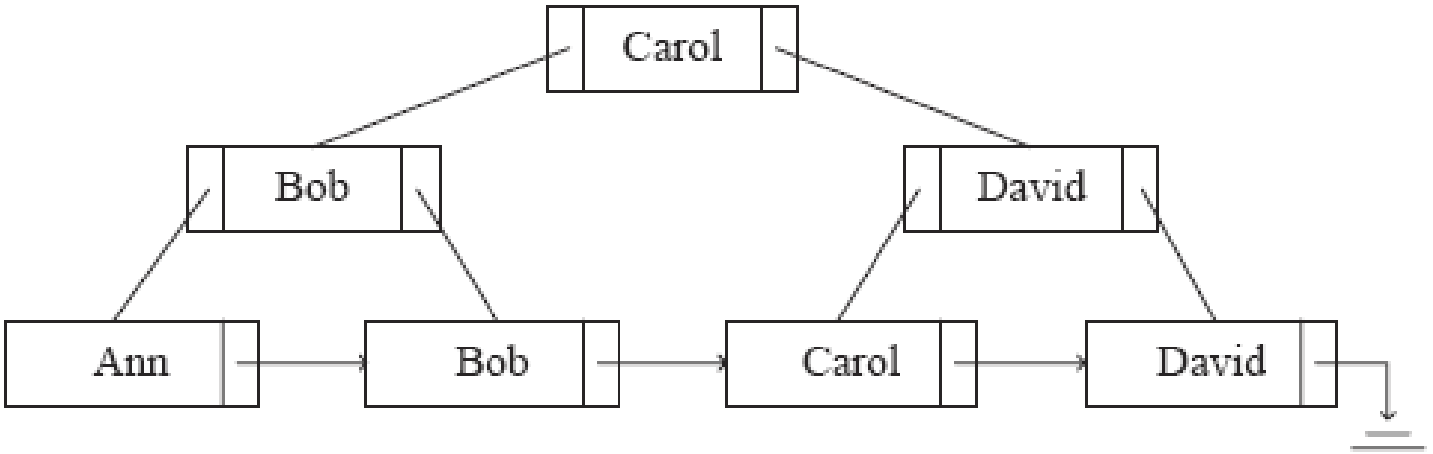
شاخص مبتنی بر درخت B+

- یکی از روش های شاخص دهی استفاده از ساختار داده ای درخت B+ است.
- در درخت B+، گره های داخلی به طور مستقیم به چندتایی ها در پایگاه داده اشاره نمی کنند بلکه به سایرگره ها در ساختار اشاره می نمایند.
- گره های برگ مستقیماً به چندتایی هایی در پایگاه داده با مقادیر مشخص برای صفت شاخص اشاره می کنند
- درخت B+ می تواند برای هر صفت A_{ij} در شمای R_i که در شروط پرس و جو ظاهر می شود، تعریف شود.
- شاخص توسط کارخواه روی مقادیر رمز نشده صفت ساخته شده و سپس به شکل رمز شده روی کارگزار همراه با پایگاه داده رمز شده ذخیره می گردد.
- ساختار درخت B+ به جدولی با دو صفت شناسه گره و محتوای گره تبدیل می شود. این جدول برای هر گره یک ردیف دارد.

مثال

Id	VertexContent
1	2, Carol, 3
2	4, Bob, 5
3	6, David, 7
4	Ann, 5, 1, 5
5	Bob, 6, 2, 3
6	Carol, 7, 4
7	David, NIL, 6

Id	C
1	gtem945/*c
2	8dq59wq*d’
3	ue63/)¡w
4	8/*5sym,p
5	mw39wio[
6	=wco21!ps
7	oieb5(p8*



روش ارزیابی پرس و جو در درخت B+

- فرض کنید، کارخواه اجرای پرس‌وجویی با شرط $A > v$ را درخواست کرده که v یک مقدار ثابت است. کارخواه باید درخت B ذخیره شده روی کارگزار را برای یافتن محل v، پیمایش کند.
  - کارخواه درخواستی برای دریافت ریشه‌ی درخت B انجام می‌دهد.(ریشه‌ی درخت، ردیفِ با شماره‌ی ۱ است.) این ردیف به کارخواه ارسال و در آنجا رمزگشایی و پردازش می‌شود.
  - با توجه به مقدار v، گره‌ی بعدی که باید پیمایش شود انتخاب شده و درخواستی مبنی بر ارسال آن گره به کارگزار فرستاده می‌شود. این روند تا پیدا کردن برگ حاوی v ادامه پیدا می‌کند.
  - پس از پیدا شدن v، تمام برگ‌های بعد از آن به سمت کارخواه ارسال می‌شوند. این گره‌ها در سمت کارخواه رمزگشایی شده و شماره‌ی رکوردهای مورد نظر پیدا شده به سمت کارگزار ارسال می‌شود.

شاخص مبتنی بر درخت +B- ویژگی ها

- درخت +B در پاسخ به پرس و جو، چندتایی اضافی به سمت کارخواه نمی فرستد.
- هزینه ارزیابی شرایط پرس و جو برای کارخواه نسبت به روش های مبتنی بر باکت و توابع درهم ساز بسیار بیشتر است.
  - به همین دلیل معمولاً درخت +B را در کنار یکی از روش های شاخص دهی مبتنی بر باکت یا توابع در هم ساز بکار برده و از درخت +B تنها در هنگام ارزیابی بازه ها در پرس و جو ها استفاده می نمایند.
- با توجه به این که محتوای درخت +B در سمت کارگزار رمزنگاری شده است، این روش در برابر حملات استنتاجی مقاوم است.

جستجوی مبتنی بر شاخص در داده های رشته ای

- روشی مبتنی بر شاخص برای پرس و جو روی داده های رشته ای توسط Wang ارائه شده است (۲۰۰۴).
- با استفاده از روش وانگ، می‌توان به جستجوی الگوهای دلخواه در داده‌های رمزشده پرداخت.
- ریزدانگی رمزنگاری در روش Wang، در سطح فیلد است.

مراحل رمزنگاری

1. به هر رشته S که از n کاراکتر $c_1c_2...c_n$ تشکیل شده است، مقدار شاخصی به طول m بیت اختصاص داده می‌شود.
2. هر رشته‌ی $S = c_1c_2...c_n$ به رشته‌ی $S' = s_1s_2...s_{2n-2}$ که در آن $s_i = c_ic_{i+1}$ است، بسط داده می‌شود.
3. با استفاده از تابع درهم‌ساز h، مقداری بین 0 تا m برای هر کدام از s_i ها به دست می‌آید.
4. اگر مقدار $h(s_i)$ برابر با k باشد. در آن صورت بیت شماره kام در شاخص مربوطه برابر با یک می‌شود.

مثال

- S1= (abcehklst)
- m=16
- تابعی در هم ساز دوتایی های ab، bc، ...، و st را به عددی بین ۰ تا ۱۵ نگاشت می کند.
- S2=Index(abcehklst)= $(0010100010101001)_2$
- هشت رشته ۲ تایی وجود دارد در حالیکه ۶ بیت یک در S2 دیده می شود. یعنی برخی از کاراکترهای دوتایی به یک مقدار توسط h نگاشت می. شوند

روش جستجو

1. مقدار رشته‌ای موجود در الگوی رشته ای مورد پرس و جو بسط داده می‌شود و مقادیر اختصاص داده شده توسط تابع h به هر دو حرف متوالی آن به دست می‌آید.
   - پرس و جو روی مقدار رشته ای به پرس و جو روی شاخص (رشته بیتی) تبدیل می شود.
2. مقادیر به دست آمده، مکان بیت‌هایی از مقادیر شاخص را که باید 1 باشند تا شرط پرس‌وجو ارضاء شود، مشخص می‌کند. این مقادیر به سمت کارگزار ارسال می‌شوند.
3. کارگزار به ازای هر مقدار شاخص، مقادیر بیت‌های آن در مکان‌های ارسال شده را بررسی می‌کند، اگر همه دارای مقدار 1 بودند، مقدار شاخص را به عنوان یکی از جواب‌ها به سمت کارخواه برمی گرداند.
   - در جواب های ارسال شده ممکن است نتایج اضافی وجود داشته باشد که باید توسط کارخواه مجدداً پردازش شده تا جواب دقیق به دست آید.

حالت های مختلف پرس و جوی رشته ای

1. تطبیق دقیق
   - شرط attribute=value روی جدول رمز نشده تبدیل می شود به a^S = index (value) که a^S مقدار رمز شده attribute در سمت کارگزار است.
2. تطبیق الگویی
   - شرط attribute LIKE value روی جدول رمز نشده:

$(a^s)_{H(c1c2)}=1$ AND … AND $(a^s)_{H(ck-1ck)}=1$) a like c0c1…ck => (($(a^s)_{H(c0c1)}=1$ AND

3. پرس و جوی با شرایط بولی

(attribute=value 1) OR (attribute=value 2), (attribute like value 1) AND (attribute like value 2), (attribute like value 1) AND NOT (attribute like value 2)

- از حالت اول و دوم قابل استنتاج است.

مثال

employee

eid	did	age	Sex
021021	Chessbasketball	24	M
021094	basketballcook	30	F
021095	Languageschat	26	M
021096	programnetwork	21	M

employee^E

eid	did^E	age	sex	did^s
021021	1001010110010010011011...	24	M	1011001011001011
021094	1001111001100001101011...	30	F	1001100001101011
021095	0110101101000111001011...	26	M	0110100011100100
021096	1111100011101011100111...	21	M	0001110101110010

select eid, age from employee where did like ‘chess’

Transforms to

select * from employee^E where

(did$^s_{H(ch)}$=1) *and* (did$^s_{H(he)}$=1) *and* (did$^s_{H(es)}$=1) *and* (did$^s_{H(ss)}$ =1)

تحلیل روش

- در این روش، در صورتی که طول m بزرگ انتخاب شود، حمله متن اصلی معلوم محتمل است.
  - هر زوج کاراکتر به مقدار یکتایی در شاخص منتسب می شوند و امکان تحلیل فرکانسی و شکستن شاخص وجود دارد.
- در صورت کوچک بودن m، تعداد زیادی داده اشتباه به سمت کارخواه ارسال خواهد شد.
- در جدول رمز شده فیلد اضافه ای برای شاخص باید در نظر گرفته شود که حجم آن به اندازه شاخص (m) وابسته است.
  - در صورتی که n فیلد نیاز به رمز شدن داشته و طول شاخص به طور متوسط m بیت باشد، میزان فضای اضافی مورد نیاز m*n بیت برای شاخص است.

جمع بندی روش ها ی مبتنی بر شاخص

- تحقیقات زیادی روی این گونه روش ها صورت گرفته و دارای زمینه تئوریک قوی می باشند. حتی دارای جبر رابطه ای مخصوص می باشند.
- بیشتر فرایند اجرای پرس و جو بر روی کارگزار متمرکز است.
- هزینه ذخیره سازی تقریباً دو برابر ذخیره سازی معمولی است. زیرا غیر از مقدار رمز شده ذخیره، مقدار شاخص نیز ذخیره می گردد.
- امکان استنتاج و افشای اطلاعات وجود دارد که میزان آن وابسته به تعریف شاخص است.
  - Evdokimov و همکارانش اثبات کرده اند که تقریباً تمام روش های مبتنی بر شاخص ذاتاً امن نیستند. به ویژه روشهایی که در پاسخ به پرس وجو چندتایی های اضافی تولید نمی کنند، در معرض تهدید افشای اطلاعات قرار دارند.
- پرس و جوهای شامل پیوند در این روش قابل اعمال است.
- پرس و جوهای تساوی، و الگویی برای داده های عددی و رشته ای قابل انجام است.
- امکان اجرای پرس و جوهای بازه ای بسته به تعریف شاخص دارد.
- امکان اجرای پرس و جوهای شامل توابع تجمعی وجود ندارد.
- چون با تغییر توزیع داده ها نیاز به دسته بندی مجدد داده ها برای تعریف شاخص وجود دارد، این روش بیشتر برای داده های فقط خواندنی مناسب است.

روش های مبتنی بر کاربرد توابع همریخت اختفایی

- توابع همریخت اختفایی نوعی توابع برای رمزنگاری است.
- در روش رمزنگاری همریخت اختفایی (privacy homomorphism)، حاصل انجام یک عمل روی داده‌های رمزشده، معادل رمزشده حاصل عملی دیگر روی داده‌های اصلی است.
- $\beta(x^E , y^E) = (\alpha(x,y))^E$
  - $E(x) . E(y) = E(x+y)$
- بوسیله‌ی توابع رمزنگاری همریخت اختفائی می‌توان برخی از عملیات مانند جمع و ضرب را به طور مستقیم روی داده‌های رمزشده انجام داد.
- فیلدهایی را که روی آنها پرس و جوی تجمعی اجرا می شود، با استفاده از رمزنگاری همریخت رمز می کنند.

روش هاسیگموس

- هاسیگموس با استفاده از یک تابع رمزنگاری همریخت اختفائی که دو خاصیت جمعی و ضربی را پشتیبانی می کند، توابع جمع و ضرب را روی داده‌های رمزشده اجرا کرد.
- خواص تابع رمزنگاری همریخت استفاده شده:
  - کلید رمزنگاری K=(p, q) که p و q به وسیله کاربر تعیین می شوند.
  - $n = p.q$ که به کارگزار داده می شود
  - $qq^{-1} = 1 \pmod p$
  - $pp^{-1} = 1 \pmod q$
  - $E_k(a \bmod p, a \bmod q)$
  - $$\mathcal{D}_k(c_1, c_2) = (c_1 \mod p)qq^{-1} + (c_2 \mod q)pp^{-1} \pmod n$$

مثال

- $p = 5 , q = 7$
- $n = p.q = 35 \quad k = (5, 7)$
- می خواهیم $a_1 = 8$ و $a_2 = 12$ را جمع کنیم
- $E(a_1) = (3,1)$
- $E(a_2) = (2,5)$
- $E(a1) + E(a2) = (3+2, 1+5) = (5, 6)$
- $D(5, 6) = ۵ + qq^{-1}*۶ \pmod{35} pp^{-1}*$

$= (5*7*3 + 6*5*3) \pmod{35} = 195 \pmod{35} = 20$

مشکل استنتاج

- استفاده از تابع رمزنگاری همریخت اختفائیِ معرفی شده کاملاً امن نیست و کارگزار می‌تواند مقدار اصلی برخی از مقادیر رمزشده را به دست آورد.
  - فرض کنید، x و y دو عدد باشند که به صورت (x_p,x_q) و (y_p,y_q) رمز شده‌اند. اگر $z=x.y$باشد، رابطه‌ی $(z_p,z_q)=(x_p,x_q).(y_p,y_q)$ نیز بین مقادیر رمزشده این متغیرها وجود دارد. در چنین حالتی کارگزار می‌تواند مقادیر اصلی این متغیرها را به دست آورد. کارگزار شروع به جمع کردن (x_p, x_q) با خودش می‌کند و این کار را آنقدر ادامه می‌دهد تا نتیجه آن با (z_p, z_q) برابر شود. تعداد این جمع کردن‌ها برابر با عدد y است.

$$(z_p,z_q)=\underbrace{(x_p,x_q)+(x_p,x_q)+\cdots+(x_p,x_q)}_{y}$$

- هاسیگموس با اضافه کردن نویز تصادفی به مقادیر رمزشده، مشکل را حل می کند. $R(x)$ یک تابع شبه تصادفی است.
- نویز در هنگام رمزگشایی از داده‌ها حذف می‌شود. $E(x)=(x\ (mod\ p)+R(x).p,\quad x.(mod\ q)+R(x).q)$

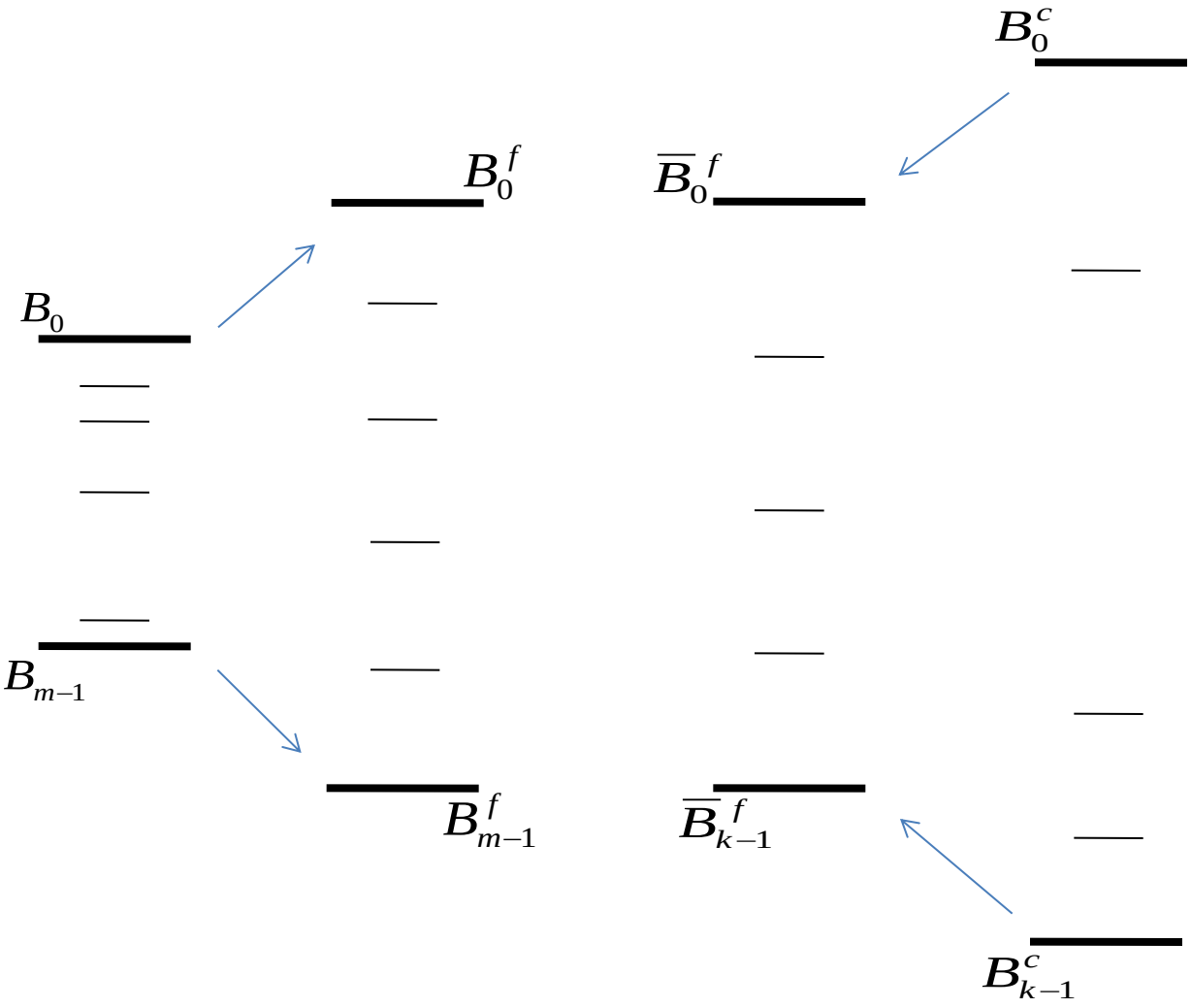
روش های رمزنگاری مبتنی بر حفظ ترتیب

- در روش های رمزنگاری با حفظ ترتیب (Order Preserving) داده ها به گونه ای رمز می شوند که ترتیب داده ها پس از رمز شدن با ترتیب داده های اصلی یکسان باشد.
- این نوع رمزنگاری برای اجرای پرس و جو های بازه ای مناسب است.
- آگراوال برای رمزنگاری داده های عددی و اجرای پرس و جوهای بازه ای روی آنها از این روش استفاده کرده است.
- فرض کنید، داده‌های اصلی دارای توزیع اولیه A هستند. این داده‌ها به نحوی رمز می‌شوند، که علاوه بر حفظ ترتیب از توزیع دلخواه A'تبعیت کنند.
  - ابتدا داده‌ها به توزیع یکنواخت f نگاشت شده و سپس توزیع f به توزیع هدف A' نگاشت می‌شود.

مراحل رمزنگاری

- OPES در سه مرحله کار می‌کند.(Order Preserving Encryption Scheme)

1. مدل کردن: داده‌های اصلی به تعدادی دسته تقسیم می‌شوند و توزیع داده‌های هر دسته مدل می‌شود.
   - روش هایی برای تعیین تعداد دسته ها و طول هر دسته وجود دارد.
   - افزایش دسته‌ها منجر به افزایش هزینه‌های مدل می‌شود.
2. مسطح کردن: داده‌های اصلی P به داده‌های مسطح Fبه قسمی تبدیل شود که مقادیر F دارای توزیع یکنواخت باشند.
   - در مرحله‌ی مسطح کردن برای هر دسته، یک تابع نگاشت M ایجاد می‌شود. تابع M هر دسته را به دسته‌ای با توزیع یکنواخت نگاشت می‌کند.
   - دو مقدار متفاوت در داده های اصلی همیشه به دو مقدار متفاوت از فضای مسطح شده نگاشت شوند.
3. تغییر: داده‌های مسطح F به داده‌های رمزشده C تبدیل می‌شود، به قسمی که مقادیر C دارای توزیع نهایی که برای داده‌های رمزشده در نظر گرفته شده بود، باشند.


B_0^c
$\overline{B}_0^f$
B_0^f
B_0
B_{m-1}
$\overline{B}_{k-1}^f$
B_{m-1}^f
B_{k-1}^c

ویژگی های روش رمزنگاری با حفظ ترتیب

- اجرای پرس و جوها در این روش چندتایی های اضافی به سمت کارخواه ارسال نمی کند.
- امنیت در این روش زمانی تامین می شود که کارگزار/حمله کننده اطلاعاتی در مورد پایگاه داده اصلی و یا دامنه صفت ها نداشته باشد.
  - این روش در مقابل حمله متن اصلی معلوم آسیب پذیر است.