جلسه هفتم

فایل با ساختار ترتیبی شاخص دار Indexed Sequential:

شاخص ها بر مبنای کلیدها و آدرس فیلدها ساخته می شوند. شاخص باعث بالا رفتن سرعت دستیابی می گردد و تکنیک شاخص بندی در اکثر نرم افزارهای امروزی استفاده می شود و جزو شیوه های دستیابی تصادفی به حساب می آید و شاخص هایی که در این جا بررسی می شود، از نوع شاخص ساده هستند.

مثال: فایل ترتیبی دانشجویان در زیر برحسب شماره دانشجویی مرتب شده می باشد که در کنار این فایل ترتیبی یک فایل ایندکس (شاخص) برحسب کلید اصلی (شماره دانشجویی) و یک فایل ایندکس برحسب معدل ترسیم شده است.

فایل ترتیبی

شماره رکورد	معدل	نام پدر	نام	شماره دانشجویی
۱	۱۷	سعید	علی	۳۹۲۵
۲	۱۹	مجید	حسن	۴۷۱۳
۳	۱۶	شاهین	امیر	۵۴۱۷
۴	۱۵	سهیل	جواد	۷۳۵۴
...	...	...	...	...

فایل ایندکس معدل

شماره رکورد	معدل
۴	۱۵
۳	۱۶
۲	۱۷
۱	۱۹
...	...

فایل ایندکس اولیه

شماره رکورد	شماره دانشجویی
۱	۳۹۲۵
۲	۴۷۱۳
۳	۵۴۱۷
۴	۷۳۵۴
...	...

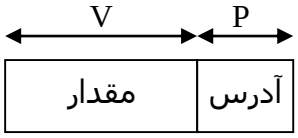
حال اگر مثلا بخواهیم مشخصات دانشجویی با معدل ۱۹ را ببینیم کافی است ابتدا در فایل کوچک ایندکس معدل، با روش باینری ستون سمت چپی را به دنبال عدد ۱۹ جست و جو کنیم بدین ترتیب متوجه خواهیم شد که مشخصات این دانشجو در سطر ۲ فایل ترتیبی اصلی قرار دارد لذا به سرعت بر سر رکورد ۲ فایل

ا صلی رفته و اطلاعات مورد نیاز را می خوانیم. بدون این فایل کمکی شاخص مجبور بودیم با جست و جوی خطی در فایل اصلی آن را پیدا کنیم که کاری زمان گیر بود.

در مثال ساده ی فوق تعداد سطرهای فایل شاخص برابر سطرهای فایل اصلی می باشد ولی تعداد ستون ها آن تنها ۲ فیلد ا ست. بدین دلیل فایل شاخص به مراتب کوچک تر از فایل ا صلی بوده و جست و جو در آن سریع تر صورت می گیرد. حتی در صورتی که فایل ایندکس خیلی کوچک باشد می توان آن را در حافظه ی اصلی نگهداری کرد و بدین ترتیب سرعت جست و جو افزایش بسیار زیادی می یابد.

اگر در فایل ایندکس صفت خاصه ی شاخص، کلید اصلی باشد به آن شاخص اولیه یا اصلی می گویند (Primary Index). و در صورتی که فایل ایندکس بر اساس فیلدی غیر از کلید اصلی ساخته شود به آن شاخص ثانویه گویند (Secondary Index).

پس فایل شاخص مجموعه ای از تعدادی مدخل (Entry) می باشد که به فرم کلی زیر:



فیلد آدرس به طول P بایت حاوی یک نشانه گر به یک یا گروهی از رکوردهاست. در فایل داده ای اصلی فیلد مقدار به طول V بایت شامل صفت خاصه ای یا ترکیبی از صفات خاصه است که ایندکس بر اساس آن ساخته شده ا ست بنابراین طول هر رکورد فایل شاخص برابر V+P بایت ا ست. به هر نقطه از فایل داده ای اصلی که از مدخل شاخص به آن نشانه گر وجود دارد را لنگرگاه یا Anchor Point گویند.

اگر هر مدخل فایل شاخص به یک رکورد ا شاره کند، شاخص را متراکم (Dense Index) گویند و اگر به گروهی از رکوردها مثلا یک بلاک اشاره کند، شاخص را غیر متراکم (Non Dense Index) گویند.

در شاخص غیر متراکم فایل اصلی داده ای باید بر اساس فیلد متناظر شاخص مرتب شده باشد تا رکوردها را بتوان دسته بندی کرد ولی در شاخص متراکم لزومی نیست که فایل داده ای از قبل مرتب باشد. فایل داده‌ای و فایل شاخص می توانند بلاک بندی شده با شند یا نشده با شند. در حالت بلاک بندی شده اغلب

اندازه ی بلاک فایل شاخص و بلاک فایل داده ای یکسان ا ست. در شاخص نا متراکم مقدار موجود در فیلد داده هر مدخل میتواند کوچک ترین یا بزرگ ترین مقدار در هر گروه باشد.

تعریف ظرفیت نشانه روی شاخص (Index fdnout):

فایل شاخص نیز مثل فایل داده ای بلاک بندی شده ا ست. تعداد مدخل های یک بلاک شاخص را ظرفیت نشـــانه روی آن می گویند. در واقع همان فاکتور بلاک بندی اســـت برای بلاک شـــاخص و با پارامتر y آن را نمایش می دهند.

$$y = \left\lfloor \frac{\text{اندازه بلاک شاخص}}{\text{طول مدخل شاخص}} \right\rfloor \rightarrow y = \left\lfloor \frac{B}{V+P} \right\rfloor$$

مثال: اگر طول بلاک ۲۰۰۰ بایت، اندازه ی صفت خا صه ی شاخص V برابر ۱۴ بایت و اندازه ی ا شاره گر شاخص P برابر ۶ بایت باشد، ظرفیت نشانه روی هر بلاک شاخص چقدر است؟

$$y = \left\lfloor \frac{2000}{14+6} = 100 \right\rfloor$$

یعنی هر بلاک شـــاخص دارای ۱۰۰ ســـطر یا مدخل اســـت و هر مدخل اشـــاره گری به رکورد یا گروهی از رکوردها در فایل داده ای اصلی می باشد و با توجه به مفروضات زیر ساختار شاخص برای این فایل چیست؟

$n = 10^6$, R=200 بایت , B=2000 بایت

در فایل اصلی ۱۰ رکورد جای میگرد

$$\frac{B}{R} = BF = \frac{2000}{200} = 10$$

$$b = \frac{n}{BF} = \frac{10^6}{10} = 10^5$$

حافظه ی مصرفی برای سطح اول شاخص $SI_1 = 10^5 \times 20$

تعداد بلاک های سطح اول $b_1 = \frac{10^5}{100} = 1000$

حافظه ی مصرفی زیاد است لذا سطح دوم شاخص را ایجاد می کنیم

برای نگه داری در حافظه ی اصلی زیاد است

$SI_2 = 1000 \times 20 = 20000$

تعداد بلاک ها در سطح دوم $b_2 = \frac{1000}{100} = 10$

پس ساختار شاخص را در سه سطح ایجاد می کنیم $SI_3 = 10 \times 20 = 200$

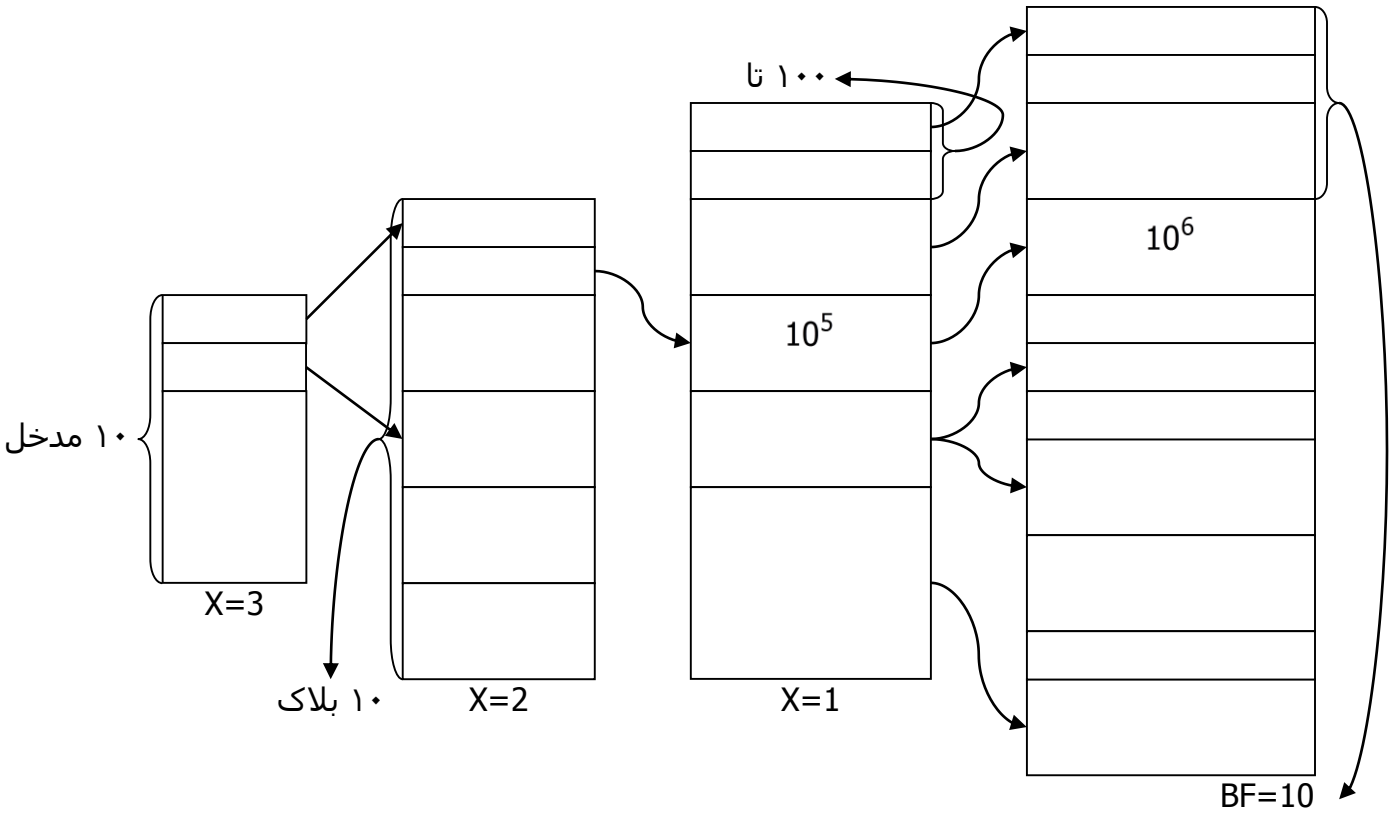
بلاک های سطح سوم $b_3 = \frac{10}{100} = 0.1$

و تعداد سطوح شاخص از رابطه ی زیر به دست می آید:

$$x = \left\lceil \log_y \left(\frac{n}{BF} \right) \right\rceil = \lceil \log_y b \rceil = \lceil \log_{100} 10^5 \rceil = 3$$

هرچه تعداد سطوح بیشتر باشد دفعات دستیابی برای واکشی رکورد بیشتر خواهد بود.

اینقدر ادامه می دهیم تا حداقل به یک بلاک برسیم و این کار را موقعی انجام می دهیم که فایل دارد ساخته می شود و شاخص هایش را می سازیم.


۱۰۰ تا
10^6
10^5
۱۰ مدخل
X=3
۱۰ بلاک
X=2
X=1
BF=10