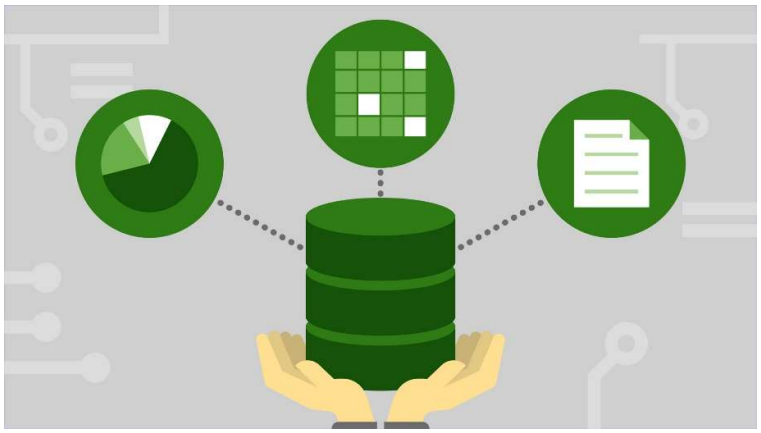
به نام خدا

دانشگاه تهران

پردیس دانشکده‌های فنی

دانشکده برق و کامپیوتر



آزمایشگاه پایگاه‌داده

دستورکار شماره ۹

مهلت تحویل :

۱۴۰۰/۰۳/۲۰

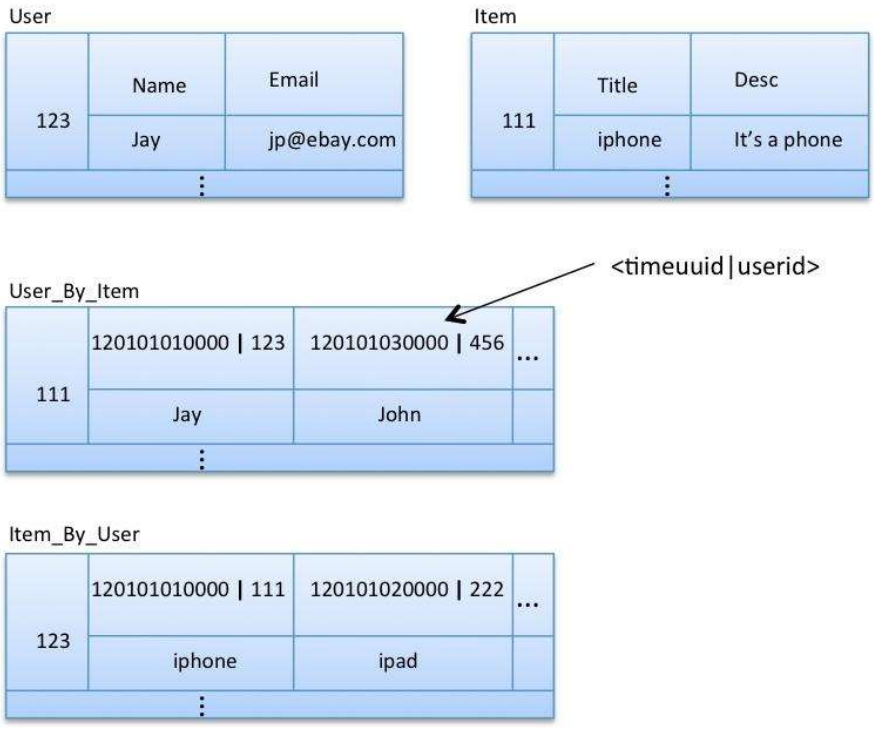
مجتبی بنائی

دستور کار شماره ۹ - کار با کاساندرا

دیتابیس‌های سطرگسترده مانند کاساندرا یا **HBase** برای ذخیره داده‌هایی ساخته شده‌اند که ساختار *کلید/مقدار* دارند به نحوی که با داشتن *کلید*، بتوان تمام اطلاعات موجود در بخش *مقدار* را با سرعت بسیار بالا خواند و پردازش کرد.

فرض کنید می‌خواهیم اطلاعات امتیازهای داده شده به یک فیلم را ذخیره کنیم. در اینجا، *کلید* همان نام یا شناسه فیلم (**RowKey**) و *مقدار*، امتیازهای مختلفی است که کاربران به آن فیلم داده‌اند . یعنی خود بخش «*مقدار*»، حاوی ستون‌ها یا بخش‌هایی (**Column Family**) است که تعداد نامشخصی دارند. یک فیلم ممکن است تنها یک نفر امتیازدهنده داشته باشد و فیلمی دیگر، هزاران امتیاز داشته باشد(جدول **ratings** که در آن،کلید هر سطر آن، نام فیلم و ستون‌های آن، هزاران امتیاز–شامل شناسه امتیاز دهنده، زمان، امتیاز– خواهد بود). **به همین دلیل به این دیتابیس‌ها، Wide Row یا Column Family می‌گوییم.**

از طرفی اگر بخواهیم بدانیم یک کاربر چه فیلم‌هایی را لایک کرده است و یا به چه فیلم‌هایی امتیاز داده است، یک جدول **UserLikes** و **UserRatings** در نظر میگیریم که در هر دوی آنها، کلید جدول، کدکاربر و ستون‌های آن ، نام فیلم‌ها خواهند بود. **بنابراین در این نوع از دیتابیس‌ها، جدول طراحی نمی کنیم بلکه دنبال یافتن کلید/مقدارهایی هستیم که هر کدام بتواند به یک کوئری مورد نیاز ما پاسخ دهد.** (هر چند ممکن است برخی اطلاعات مانند اطلاعات خود یوزر و فیلم ظاهراً به صورت جدول عادی ذخیره شوند - این جداول هم پشت صحنه با قالب کلید مقدار ذخیره می‌شوند– اما سایر جداول، ساختاری کاملا مطابق به قالب کلید/مقدار خواهند داشت– شکل زیر)



در طراحی دیتابیس‌های سطر گسترده ، باید دید رابطه‌ای را کنار بگذارید و بسته به نیاز اطلاعاتی و جستجوهایی که انجام خواهید داد، به طراحی جداول بپردازید. نگران افزونگی و تکرار داده‌ها نباشید چون برای بالابردن سرعت جستجو در بین میلیون‌ها رکوردی که در بین ده‌ها نود شبکه پخش شده‌اند، مجبوریم فضای دیسک را بیشتر از حالت نرمال، مصرف کنیم.

دقت کنید که معماری کاساندرا به گونه ای طراحی شده است که بتوان از بین میلیاردها رکورد،داده‌های مورد نیاز کاربر را با داشتن کلید سطر، به راحتی پیدا کرد . **بنابراین برای تضمین سرعت بالا در باز یابی داده‌ها، امکاناتی مانند جو ین یا اتصال جداول، مرتب سازی و مانند آنرا در این دیتابیس ندار یم (یا به شکل بسیار محدود).**

توصیه می‌کنیم قبل از شروع کار با بخش از تمرین ، دو مقاله **Cassandra Data Modeling Best Practices** با عنوان **ebay** را که به صورت عملی و با ذکر یک مثال، به بررسی نحوه مدلسازی داده‌ها در کاساندرا پرداخته است را حتما مطالعه کنید. در ادامه، توضیحاتی مختصر راجع به این دیتابیس و مفاهیم پایه آن ذکر می کنیم و سپس به بیان خود تمرین این بخش خواهیم پرداخت.

پایگاه داده سطر گسترده کاساندرا یکی از محبوب‌ترین دیتابیس‌های **NoSQL** است. در کاساندرا جداول در **keyspace**ها(معادل دیتابیس در بانک‌های اطلاعاتی رابطه‌ای) قرار می‌گیرند و هر نود می‌تواند شامل یک یا چند **keyspace** باشد و هر **keyspace** دارای استراتژی تکرار(**Replication**) و توزیع(**Partitioning**) مخصوص به خودش است. سپس با تعریف جداول با ستون‌های مشخص می‌توان اطلاعات را در سطرها ذخیره کرد.

در کاساندرا نظیر دیتابیس‌های دیگر هر سطر دارای یک کلید است اما مفهوم و کارکرد کلید در کاساندرا کمی با سایر دیتابیس‌ها متفاوت است. در کاساندرا نحوه توزیع داده‌ها بین نودها بر اساس **Partition Key** و نحوه مرتب‌سازی داده‌ها در هر پارتیشن، براساس **Clustering Key** انجام می‌شود. دقت کنید که داده‌ها در کاساندرا هنگام ذخیره سورت می‌شوند و هنگام بازیابی، نمی‌توانید دستور سورت داده‌ها را براساس فیلدی غیر از آنچه در کلاسترکی مشخص شده است بدهید. کلیدها در کاساندرا می‌تواند ساده یا ترکیبی باشند و با توجه به شرایط هر جدول می‌تواند **Partition Key** و **Clustering Key** چندمقداری داشته باشد.(برای آشنایی با این مفاهیم می‌توانید به این مقاله فارسی مراجعه کنید[1])

نکته‌ای که در کار کردن با دیتابیس‌های سطر گسترده شبیه به کاساندرا باید به آن توجه کنیم این است که تکرار داده در جداول مختلف امری طبیعی است و معمولاً نمی‌توان با طراحی یک جدول به تمام سؤالات پاسخ داد و بر اساس نیازمندی‌های سؤالات مختلف باید جدول مربوط به آنرا طراحی کنیم.

نصب و راه‌اندازی

برای کار با دیتابیس کاساندرا به **JDK-8** نیاز داریم. (دقت کنید که نسخه مناسب را نصب کنید وگرنه ممکن است به خطاهای مختلف و ناشناخته‌ای برخورید) سپس کاساندرا را با استفاده از این راهنماها نصب می‌کنیم:

windows : https://phoenixnap.com/kb/install-cassandra-on-windows

ubuntu : https://phoenixnap.com/kb/install-cassandra-on-ubuntu

و سپس در صورت نیاز درایور کاساندرا برای پایتون را نصب می‌کنیم. پس از نصب کاساندرا که با داکر هم می‌تواند راه‌اندازی شود این آموزش را انجام دهید تا برای این دستور کار، آمادگی اولیه را پیدا کنید :

https://medium.com/@aymannaitcherif/beginners-guide-to-learn-cassandra-part-2-4e8511a4838f

[1] http://yun.ir/hdww7c

دستور کار :

می خواهیم توئیت های سهام یاب را به کمک API داده شده در تمرینات قبلی،دریافت کنیم و امکانی فراهم کنیم که با داشتن یک کاربر یا یک هشتگ خاص،تمامی توئیت‌های متناظر با آن در یک بازه زمانی به ما برگشت داده شود.

جداول لازم را در کاساندرا برای این موضوع طراحی کنید (جدول کاربر / جدول هشتگ) **و به کمک درایور پایتون کاساندرا، همزمان با در یافت توئیت‌ها، آنها را در کاساندرا در این دو جدول ذخیره نمایید**. معمولا خود توئیت‌ها را در الستیک سرچ ذخیره میکنیم (به دلیل ماهیت متنی آنها) و آی دی آنها را درکاساندرا ذخیره خواهیم کرد. برای این دستور کار می‌توانید یک جدول توئیت هم در کاساندرا تعریف کنید که خود توئیت به همراه اطلاعات اصلی آن مانند متن توئیت، هشتگ، نام کاربر و زمان ارسال در آن ذخیره شود و جدول کاربر و هشتگ، تنها آی دی این توئیت را ذخیره کند. (در حالت حرفه‌ای و تجاری، این کار را انجام نمیدهیم و داده‌های اصلی در دیتابیس‌های رابطه‌ای یا الستیک سرچ و مانند آن ذخیره میشوند و کاساندرا فقط نقش یک ایندکس ثانویه برای جستجوی سریع کلیدها را برعهده دارد)

در یک برنامه دیگر، کدی بنویسید که با گرفتن نام کاربر یا هشتگ، تمام توئیت‌های متناظر با آنها را در یک بازه زمانی که آنرا هم میتوانید از کاربر در یافت کنید، نمایش دهد.

آیا می‌توانید تعداد توئیت‌ها در یک بازه خاص، تعداد توئیت‌های یک کاربر در یک بازه و تعداد توئیت‌های یک هشتگ در یک بازه را هم نمایش دهید ؟

خروجی شما در این دستور کار، دو فایل پایتون فوق به همراه توضیح مختصر روند طراحی جداول و کدهای نوشته شده خواهد بود.