**به نام خدا**

**علی عباس حقیقت**

**220797069**

**موضوع بخش دوم تمرین اول درس داده کاوی :**

**MapReduce – KNN**

همانطور که در توضیح تمرین گفته شده در ابتدای کار ما 5 عدد سرور مختلف داریم پس میشود در نظر گرفت 5 تا سرور محلی و یک سرور مستر داریم .

حال در گام اول میاییم به صورت توزیع شده در هر کدام از سرور های محلی عملیات map را انجام میدهیم یعنی دیتای خام خود را به <key , value> تبدیل میکنیم . در این محل key ما id سبد خرید(عدد int نه رقمی) و value ما id محصول است که به ترتیب از 1 تا n شماره گذاری شده است.

پس در این گام از حالتی که value را 1 بگیریم صرفنظر میکنیم زیرا یک دیتای بی ارزش را برای ما نگهداری میکند .در پایان این گام تمام داده ما تبدیل به

<basket\_id , product\_id> میشود .

سپس وارد گام بعدی یعنی sort میشویم به این معنا که تمام keyValue ها خود را بر اساس برابری کلید یعنی basket\_id در کنار هم میچینیم .(این گام هم در سرور های محلی صورت میگیرد)

سپس وارد گام بعدی میشویم که مرحل نهایی الگوریتم نگاشت کاهش یعنی reduce است . در این بخش کاری که میکنیم این است که تک تک KeyValue های با کلید یکسان را در یک فرم جدید نگه میداریم یعنی کلید ما همان basket\_id و value ما یک دیکشنری است که هریک از جفت های داخلش یک Key برابر با product\_id دیگری برابر با count دارد که همان تعداد محصولات مشابه داخل یک سبد خرید است . در این مرحله یعنی reduce اگر یک کالا از قبل داخل سبد بود مقدار count آن را میفزاییم و اگرنه یک KeyValue جدید میفزاییم. این کار با اینکه پیچیدگی کد را زیاد میکند اما از ایجاد یک sparse matrix با تعداد n خانه (با توجه به مختصات مثال یعنی آمازون میتواند در حدود 100000 باشد)جلوگیری میکند .

در گام بعدی برای پیاده سازی KNN باید تمام داده هایی را که حال پردازش شده و آماده اند را به سرور master منتقل کنیم .

حال در اینجا به پیاده سازی الگوریتم KNN میرسیم . در این مرحله با توجه به محدودیت رم و همچنین در نظرگرفتن این مسئله که شاید چند میلیارد سبد خرید داشته باشیم باید از روش sampeling با 1/100 حجم دیتای اولیه استفاده کنیم .

سپس میاییم دیتای لود شده داخل رم را با آن سبد خرید خاص مد نظر مورد مقایسه قرار میدهیم یعنی قدر مطلق حاصل کم کردن count محصولات را بدست میاوریم و اگر در یک سبد خرید یک کالایی وجود نداشت count پیش فرض 0 است و اگر آن محصول در هر دو سبد خرید نبود آن محصول را در نظر نمیگیریم . حال میاییم تمام قدر مطلق های محاسبه شده را که هر کدام حاصل مقایسه تعداد دو محصول با id یکسان در دو سبد خرید هستند ، را با هم جمع میکنیم .

حال می آییم یک k فرض میکنیم این مقدار k فاکتور های زیادی برای محاسبه دارد اما ما فرض میکنیم شرکت آمازون با توجه به سابقه خود به k با مقدار 46 رسیده است .(بیشتر بر اساس تست و سعی و خطا)

حال میاییم 46 سبد خرید را کمترین فاصله (جمع تمام قدر مطلق ها)را با سبد خرید خاص دارند میابیم .

و میایم محاسبه میکنیم از این 46 همسایه کدام گروه از افراد (مثلا افرادی که وفادارند یا نه ) وفادار به آمازون یا عدم وفاداری به آمازون بیشتر در همسایگی آن سبد خرید خاص ما هستند . حال اگر هر کدام بیشتر باشند متوجه میشویم که سبد خرید مذکور با احتمال بالا متعلق به مشتریان وفادار است یا خیر .

نکات برخی از توضیحات و فرض ها داخل فایل تکست .

فرض مهم : ما دریک فایل جداگانه مشخصا داریم که هر product\_id یا هر basket­­­­­­\_id به چه شخص یا کالایی تعلق دارد فلذا برای صرفه جویی در پراسس و حافظه از id ها به تنهایی استفاده می کنیم .

سپاس از توجه شما