## دستورالعمل های ویژه

شما باید یک README.txt اضافه کنید و هر جزئیاتی را که به آزمایش/اجرای اسکریپت شما کمک می کند اضافه کنید. باید بتوانم فایل ها را از حالت تار خارج کنم و یک driving script را از آن دایرکتوری بدون مشکل اجرا کنم.

## معرفی

در این پروژه با استفاده از ابزارهای لینوکس کارهایی شبیه به موتور ETL انجام خواهید داد. همه موارد زیر باید با یک

driving bash scrip که هر مرحله یا چندین مرحله را اجرا می کند، خودکار شوند.

از انواع ابزارها/فیلترهای مختلف لینوکس برای انجام پروژه (tr، awk، sed، bash، sort، cut و comm, wc, tail, paste, join, split) استفاده کنید. سعی کنید از یک ابزار/برنامه برای حل هر مشکل استفاده نکنید، بلکه از ابزارهای مختلف استفاده کنید. به عنوان مثال، awk یک ابزار خوب برای گزارش ها و یک کاندید خوب برای چندین مرحله دیگر است.

هر مرحله شماره گذاری شده باید به عنوان یک اسکریپت یا ابزار جداگانه در اسکریپت شما نوشته شود. به عبارت دیگر یک برنامه big piped برای انجام تمام مراحل ننویسید. هدف این است که برنامه های کوچکتری جداگانه در اسکریپت خود داشته باشید که می توانند بعداً برای سایر مشکلات ETL دوباره استفاده شوند.

اسکریپت باید خطاهای منطقی را به دام بیاندازد و پیام های خطای مفیدی را برای کمک به کاربر در حل مشکلات چاپ کند.

اسکریپت باید پیام هایی را در خروجی استاندارد چاپ کند تا نشان دهد کدام مراحل فرآیند ETL تکمیل شده است.

اگر هیچ پارامتری به اسکریپت شما ارسال نشده است، باید یک عبارت استفاده را چاپ کند که تشخیص نحوه اجرای اسکریپت را برای کاربر آسان کند. اگر نتوانم بفهمم که چگونه اسکریپت شما را اجرا کنم، نمی توانم آن را درجه بندی کنم!

اسکریپت bash باید پارامترهای زیر را بپذیرد:

- پارامترهای انتقال فایل از راه دور Remote file transfer parameters

(1) remote-server : نام سرور یا آدرس IP.

(2) remote-userid: userid برای ورود به دستگاه راه دور (فرض کنید از کلیدهای ssh استفاده می کنید تا رمز عبور لازم نباشد).

(3) remote-file: مسیر کامل به فایل راه دور در سرور راه دور.

- الزامات پارامترهای MariaDB

(4) mysql-user-id: sql user id

(5) mysql-database: sql database name

در نظر داشته باشید که برای درجه بندی این اسکریپت با فایلی متفاوت از فایل تست, تست می شود.

سعی کنید به این فکر کنید که با چه نوع خطاهایی ممکن است مواجه شوید و آنها را در اسکریپت خود مدیریت کنید.

نمونه هایی از این را می توانید در بخش Testing your script پیدا کنید

Sort - General Info

– از default sort order استفاده کنید، مگر اینکه غیر از این مشخص شده باشد.

- به داده های منبع توجه زیادی داشته باشید. داده هایی که منحصراً عددی هستند باید به این ترتیب مرتب شوند.

یادداشت ویژه در مورد فایل منبع MOCK MIX v2.1.csv.bz2

- نام فایل منبع می تواند متفاوت باشد (مطمئن شوید که نام فایل را کدگذاری نکنید)

- محل (مسیر فایل) فایل منبع می تواند متفاوت باشد (مطمئن شوید که مسیر فایل را کدگذاری سخت نکنید)

– نوع فایل منبع همیشه csv. خواهد بود

– فشرده سازی فایل csv. همیشه bz2. خواهد بود

جزئیات پروژه

آدرس IP سرور: 40.69.135.45

نام فایل منبع و مکان: home/shared/MOCK MIX v2.1.csv.bz2/

اسکریپت ETL شما (etl.sh) باید اسکریپت ها یا ابزارهای لینوکس را اجرا کند که تمام مراحل زیر را انجام می دهند:

1. فایل منبع را با استفاده از دستور scp به دایرکتوری پروژه خود منتقل کنید.

دستور scp باید متغیرهای remote-server، remote-usrid و remote-file را به عنوان آرگومان هایی برای ورودی مورد نیاز خود بپذیرد.

فایل منبع حاوی فیلدهایی است ( customer id, first name, last name, email, gender, purchase amount, credit (card, transaction id, transaction date, street, city, state, zip, phone

برای فرمت قطعی به سربرگ فایل های منبع مراجعه کنید.

فایل منبع از این نقطه به بعد به عنوان transaction file ارجاع داده می شود و باید در کد شما به این شکل نامگذاری شود.

2. transaction file را از حالت فشرده خارج کنید.

3. header record را از transaction file حذف کنید.

4. تمام متن های موجود در فایل تراکنش را به حروف کوچک تبدیل کنید.

5. فیلد "gender" می‌تواند حاوی مقادیر ««f، m»»، "male“ ,”female" ، «1»، «0»، «u»، «و...» باشد -

آنها را به صورت زیر تبدیل کنید:

"1" تا "f"

"0" تا "m".

”male” to “m“

”female” to “f“

"u" برای تمام فیلدهای دیگری که با معیارهای بالا مطابقت ندارند.

بعد از این مرحله، قسمت جنسیت فقط باید ««m، f»» یا «u» داشته باشد.

6. تمام رکوردهای transaction file را از قسمت "state" که حالت ندارند یا حاوی "NA" هستند، فیلتر کنید. این رکوردها را در یک فایل exceptions به نام extremes.csv قرار دهید.

توجه: این exceptions دیگر نباید در transaction file قرار گیرند.

7. علامت $ موجود در transaction file را از قسمت purchase amt حذف کنید.

8. مرتب سازی فایل تراکنش بر اساس customerID. فرمت transaction file نباید تغییر کند. فقط ترتیب مرتب سازی باید متفاوت باشد. transaction file نهایی باید transaction.csv نام داشته باشد.

9. یک summary file با استفاده از فایلtransaction.csv ایجاد کنید. کل total purchase را برای هر customerID جمع آوری کنید و یک فایل جدید با یک رکورد به ازای هر customerID و کل مبلغ روی همه سوابق برای آن مشتری تولید کنید. از کاما به عنوان جداکننده فیلد خود استفاده کنید.

(الف) فیلدهای این فایل باید به ترتیب زیر باشد:

1. customerID

2. state

3. zip

4. lastname

5. firstname

6. total purchase amount

(ب) فایل خلاصه را بر اساس مرتب کنید (در این مرحله مراقب باشید، کلید این کار ترتیب مرتب سازی "priority sort" است):

1. state

2. zip (descending order)

3. lastname

4. firstname

فرمت فایل را همانگونه که در مرحله 9(a) مشخص شده است نگه دارید. فقط ترتیب مرتب سازی باید متفاوت باشد. فایل خلاصه نهایی باید summary.csv نام داشته باشد.

10. دو گزارش زیر را با استفاده از فایلtransaction.csv ایجاد کنید.

(الف) Transaction Report - تعداد تراکنش ها را با state abbreviation نشان می دهد. مخفف حالت باید بزرگ باشد. این گزارش باید دارای دو عنوان در بالا باشد «گزارش توسط: [FirstName LastName]» و «Transaction Count Report ». همچنین باید برای هر ستون سرصفحه ستونی (State and Transaction Count) داشته باشید. گزارش باید بر اساس تعداد تراکنش ها به ترتیب نزولی و سپس state مرتب شود. نام این گزارش را trade.rpt بگذارید. برای فرمت دقیق فایل، به « Example Transaction Report » در بخش خروجی نمونه مراجعه کنید.

(ب) Purchase Report - کل خریدها را بر اساس جنسیت و ایالت نشان می دهد. مخفف حالت و جنسیت باید با حروف بزرگ باشد. گزارش باید دارای دو عنوان در بالا باشد «گزارش توسط:

[FirstName LastName]" "Purchase Total Report" ". همچنین باید برای هر ستون سرصفحه ستونی داشته باشید ( State, Gender, Purchase Amount). گزارش باید بر اساس تعداد کل خریدها به ترتیب نزولی، سپس بر اساس ایالت و در نهایت بر اساس جنسیت مرتب شود. توجه: قسمت Purchase Amount دارای اعداد گرد شده به نزدیکترین صدها است.

نام این گزارش را buy.rpt بگذارید.

برای فرمت دقیق فایل، «Example Purchase Report » را در بخش نمونه خروجی ببینید.

11. این یک پروژه تحقیقاتی است، بنابراین شما باید در مورد فرآیند پیاده سازی در مورد چگونگی انجام این کار تحقیق کنید. شما باید سرور mariadb را روی توزیع لینوکس خود نصب کنید.

پس از اینکه اسکریپت شما دو گزارش بالا را تولید کرد، اسکریپت شما باید رمز عبور پایگاه داده را درخواست کند. توجه: کاربر نباید قادر به دیدن رمز عبور وارد شده باشد. رمز عبور باید در متغیری ذخیره شود که بعداً در دستور mysql - password استفاده می شود. وظیفه شما بارگذاری فایل ها در mysql است.

فراموش نکنید که از پارامترهای ارسال شده برای mysql-user-id و mysql-database به عنوان آرگومان های ارائه شده به دستورات mysql خود استفاده کنید. واردات باید از رمز عبور mysql استفاده کند (از رمز عبور متنی خالی یا کدگذاری شده سخت استفاده نکنید). ممکن است بخواهید به استفاده از mysqlimport برای پیاده سازی در اسکریپت etl.sh خود نگاه کنید. همچنین ممکن است بخواهید از گزینه --local در mysqlimport استفاده کنید. طرح بندی جدول باید با چیدمان csv مطابقت داشته باشد. می توانید جداول را قبل از اجرای اسکریپت ایجاد کنید و فرض کنید جداول قبل از اجرای import وجود دارند. برای یک چالش اضافی، می توانید جداول را از این اسکریپت نیز رها کرده و ایجاد کنید.

(الف) فایل "transaction.csv" را در جدول "TRANSACTION" در پایگاه داده mysql در MySQL/MariaDB بارگیری کنید.

انواع داده MySQL برای فیلدها:

transaction date - DATE ، purchase amount - (13،2)DECIMAL ، بیشتر فیلدها، VARCHAR -

(ب) فایل "summary.csv" را در جدول "SUMMARY" در پایگاه داده mysql در MySQL/MariaDB بارگیری کنید.

انواع داده برای فیلدها:

VARCHAR برای اکثر فیلدها، DECIMAL (13،2) - purchase amount.

12. پس از اجرای اسکریپت، تمامی فایل های کاری میانی باید حذف شوند. تنها فایل‌های باقی‌مانده باید فایل‌های transact.csv، و extract.csv، summary.csv،transaction.rpt، buy.rpt باشند. در صورت خروج اسکریپت ها با خطا، این فایل ها نباید حذف شوند.

سایر موارد روبریک

_استانداردهای کدنویسی و قابلیت استفاده مجدد برای همه اسکریپت ها شامل بیانیه استفاده و سایر استانداردهای کدگذاری

_خوانایی/مستند برای همه اسکریپت ها

README: نحوه اجرای اسکریپت باید در فایل README.txt گنجانده شود.

_نظرات!!! شما باید کد خود را کامنت کنید

_هر چیز دیگری که در روبریک کدنویسی مشخص شده است.

تست اسکریپت شما

اسکریپت را با اعتبارنامه انتقال فایل بد اجرا کنید.

اسکریپت را با فایل منبع غیر موجود اجرا کنید.

فایل آزمایشی را با خطاهای زیر در برخی از فیلدها بسازید

- مقدار "value of blank " Gender و "x" (تبدیل به "u")

- مقدار "value of blank "state و "NA" (این یک استثنا خواهد بود)

- حروف مختلط در زمینه نام و خیابان

_کل خریدهای انباشته شده را تأیید کنید

_گزارش ها را تایید کنید

_فایل های تراکنش و خلاصه بارگذاری شده در MySQL را تأیید کنید

توجه داشته باشید

ممکن است بخواهید فایل را به دستگاه محلی خود برای آزمایش بومی سازی شده در صورتی که سرور غیرقابل دسترسی است یا برای آزمایش آفلاین اسکپ کنید.

همه ارزها باید در قالب 2 دسیمال واحد باشند! مثال: 33734.33

مطمئن شوید که اسکریپت شما داخل یک پوشه با نام شما روی آن است.

نمونه خروجی

این بخش شامل دو فایل خروجی می باشد. توجه: این 7 خط اول خروجی واقعی (پاسخ های صحیح) است که گزارش شما باید تولید کند. نام و نام خانوادگی خود را جایگزین مکان نگهدار ([نام شما]) کنید.

:Example Transaction Report

$ head -n 7 transaction-rpt

Report by: [Your Name]

Transaction Count Report

| State | Transaction Count |
|---|---|
| TX | 131 |
| CA | 103 |
| FL | 96 |
| NY | 60 |

:Example Transaction Report

$ head -n 7 purchase-rpt

Report by: [Your Name]

Purchase Summary Report

| State | Gender | Report |
|---|---|---|
| TX | F | 33734.33 |
| CA | F | 23911.61 |
| TX | M | 23043.64 |
| FL | M | 18846.49 |

فایل های زیر را ارسال کنید:

README.txt

etl.sh

هر دایرکتوری یا اسکریپت پشتیبانی کننده به etl.sh

فایل های زیر را ارسال نکنید، زیرا اسکریپت شما باید این فایل ها را ایجاد کند:

buy.rpt ،transaction.rpt،summary.csv ،exceptions.csv،transactions.csv،MOCK_MIX*.tar.bz2

هشدار: فقط تکالیف کدنویسی را به زبان های برنامه نویسی ارسال کنید: Bash، AWK، Sed.