ETL Pipeline

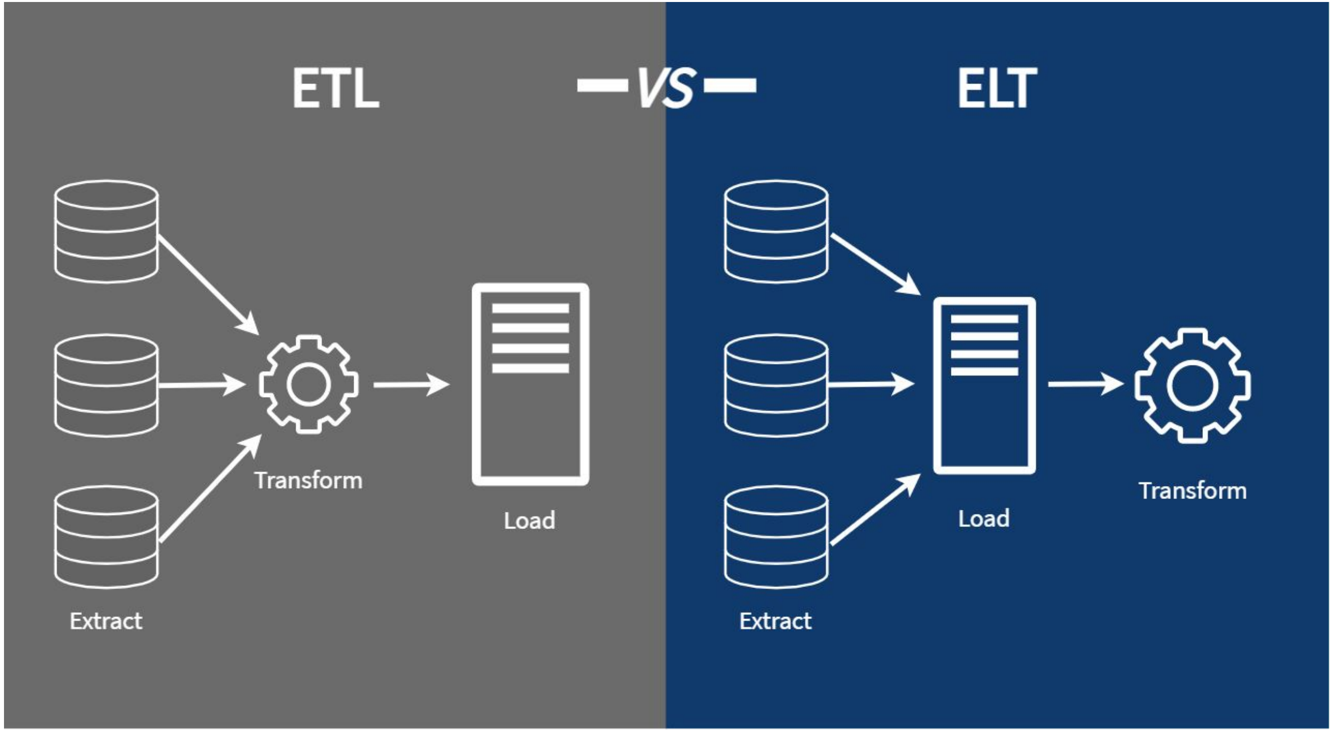
Coded Using Python & SQL on a MySQL Database

Connie Sau Chow

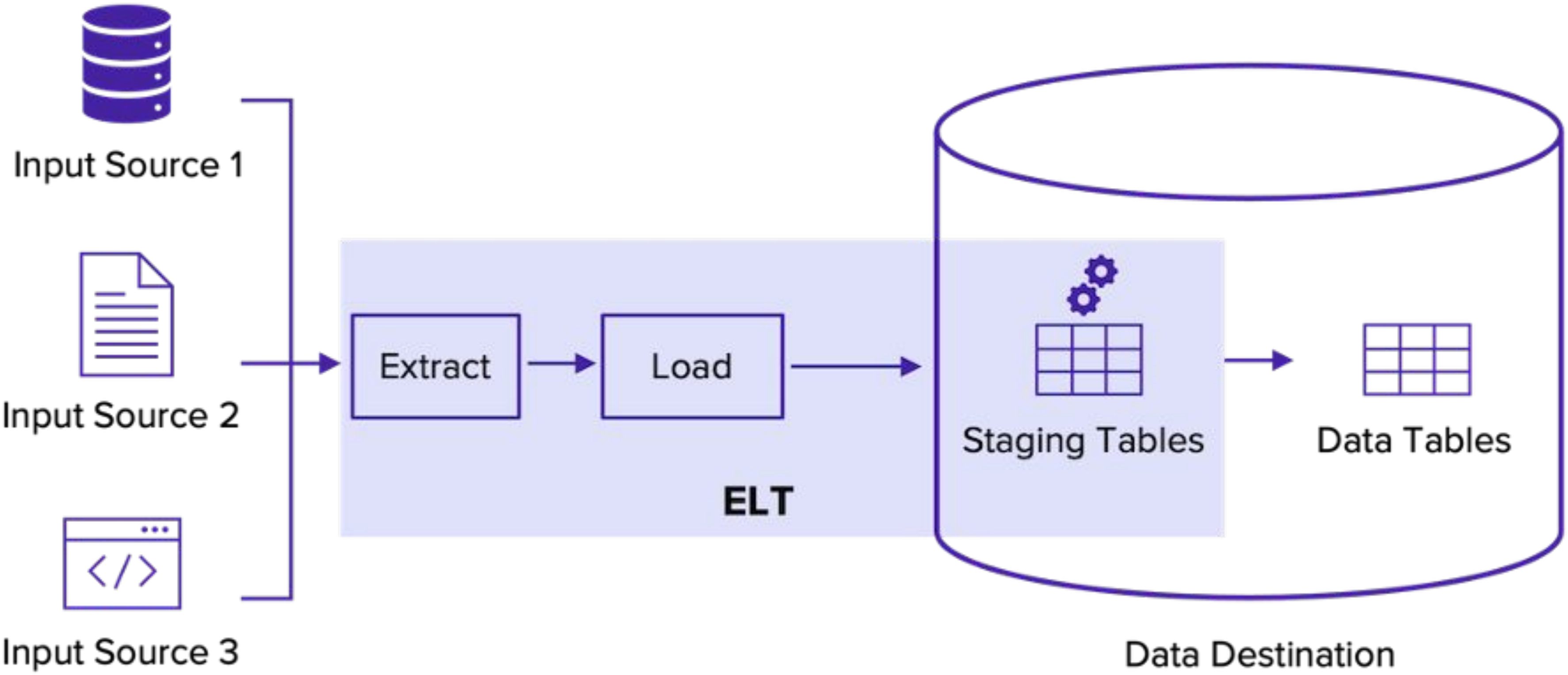
Overview

- ELT Pipeline Design
- Application Design & Architecture
- Database Design
- Application Sequence Diagram
- Limitations & Scalability
- Next Steps & Future Development Features

ETL Design Paradigm



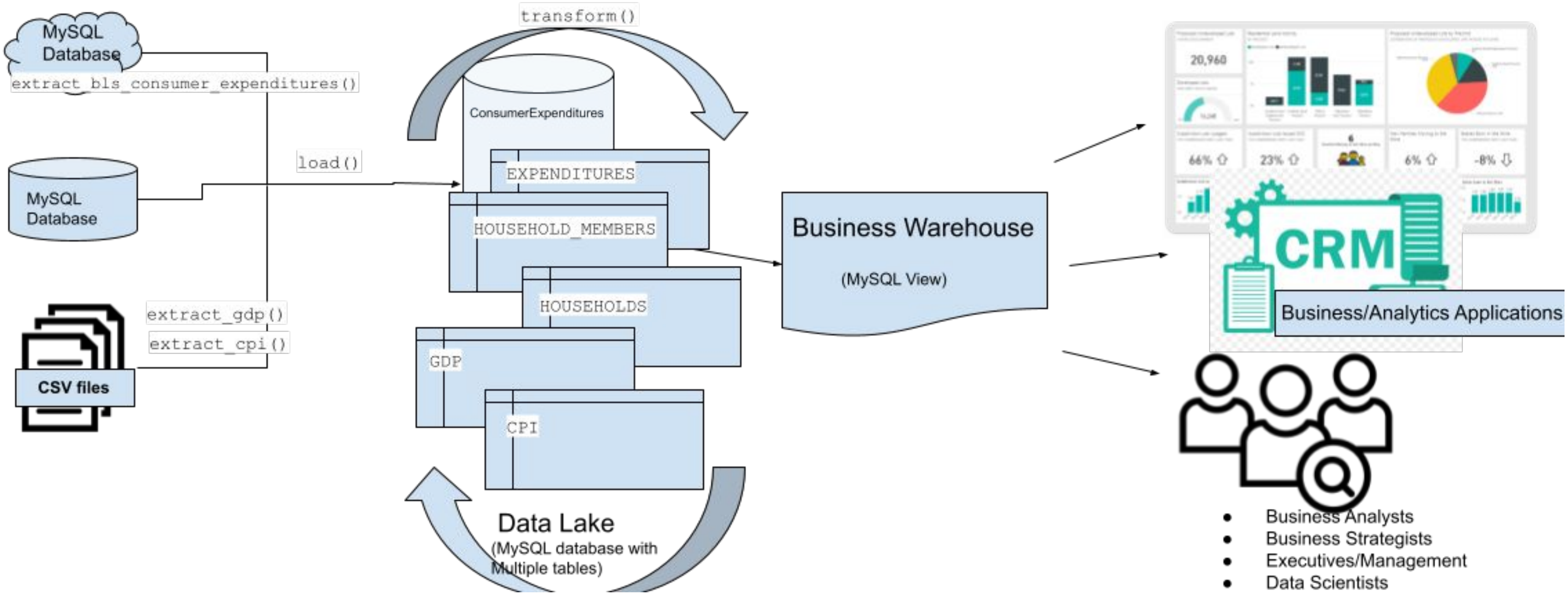
High Level Design & Architecture



Data Sources

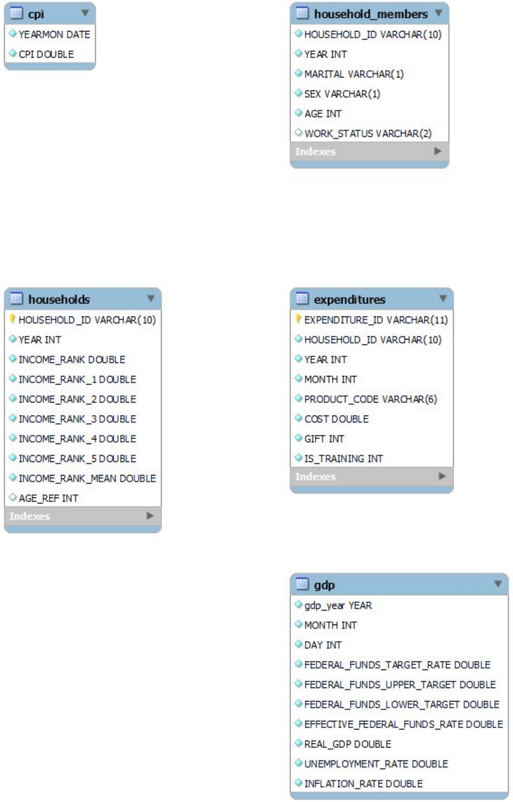
	Size	Row Count	Number of Columns
Consumer Expenditures	337.6 MB	2,047,961	10
Household Members		137,355	6
Households		56,812	11
GDP	26.5 KB	904	10
CPI	22.7 KB	1,303	2

ELT Python Application Design & Architecture



- Business Analysts
- Business Strategists
- Executives/Management
- Data Scientists

EER Diagram (Database Schema)



cpi
YEARMON DATE
CPI DOUBLE

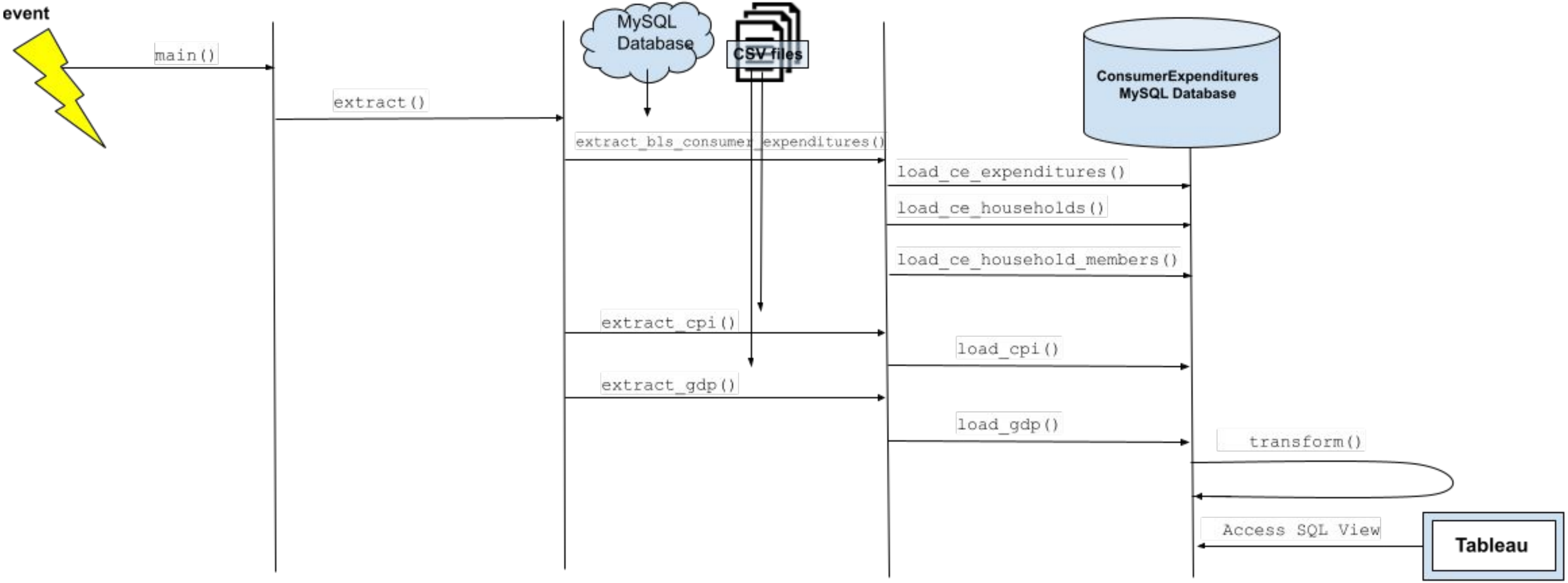
household_members
HOUSEHOLD_ID VARCHAR(10)
YEAR INT
MARITAL VARCHAR(1)
SEX VARCHAR(1)
AGE INT
WORK_STATUS VARCHAR(2)
Indexes

households
HOUSEHOLD_ID VARCHAR(10)
YEAR INT
INCOME_RANK DOUBLE
INCOME_RANK_1 DOUBLE
INCOME_RANK_2 DOUBLE
INCOME_RANK_3 DOUBLE
INCOME_RANK_4 DOUBLE
INCOME_RANK_5 DOUBLE
INCOME_RANK_MEAN DOUBLE
AGE_REF INT
Indexes

expenditures
EXPENDITURE_ID VARCHAR(11)
HOUSEHOLD_ID VARCHAR(10)
YEAR INT
MONTH INT
PRODUCT_CODE VARCHAR(6)
COST DOUBLE
GIFT INT
IS_TRAINING INT
Indexes

gdp
gdp_year YEAR
MONTH INT
DAY INT
FEDERAL_FUNDS_TARGET_RATE DOUBLE
FEDERAL_FUNDS_UPPER_TARGET DOUBLE
FEDERAL_FUNDS_LOWER_TARGET DOUBLE
EFFECTIVE_FEDERAL_FUNDS_RATE DOUBLE
REAL_GDP DOUBLE
UNEMPLOYMENT_RATE DOUBLE
INFLATION_RATE DOUBLE

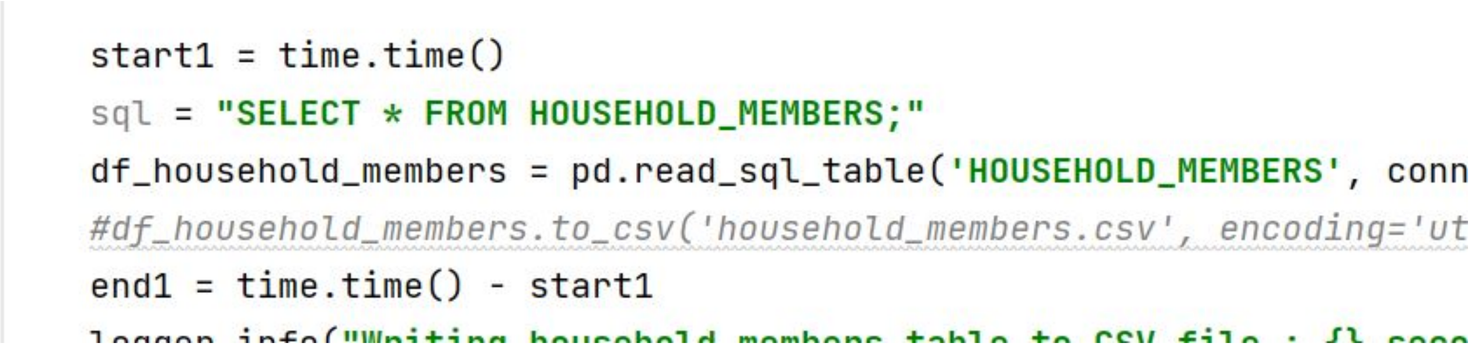
ELT Application Sequence Diagram



Limitations

Dynamic Extraction of Tables

Hardcoded Table Names and Column Types



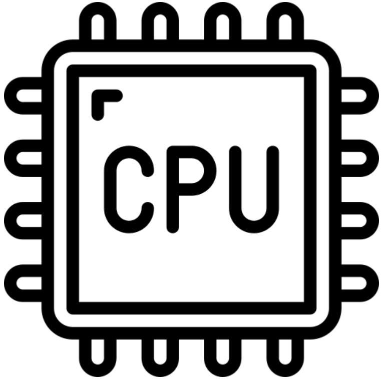
```
start1 = time.time()
sql = "SELECT * FROM HOUSEHOLD_MEMBERS;"
df_household_members = pd.read_sql_table('HOUSEHOLD_MEMBERS', conn
#df_household_members.to_csv('household_members.csv', encoding='ut
end1 = time.time() - start1
logger.info("Writing household members table to CSV file : {} seco
```

Limitations

Lacks Scalability For Larger Datasets and Data Processing Capacity

Uses Local MySQL Database

Uses SQLAlchemy libraries which may not be most efficient



pixtastock.com – 74331349

Limitations

Maintainability of Code

Needs centralized configuration file

Needs centralized logging options for different users

shutterstock.com · 714939874

Future Implementation Items

More robust ELT pipeline that can scale & integrate

- *Automatic scheduled pulls configurable per data source*
- *Dynamic extraction of external data sources*
- *New Pipeline Segment to batch extract and combine data sources before writing to database (Data Lake/Staging Area)*
- *Configurable Log Files Based on User Requirements*
- *Centralized Configuration File for Storing file locations, etc.*