**Data Warehouses and Data Analysis Techniques  
PMS 2024**

**Γεώργιος Βώβος**

**Αριθμός μητρώου: Μ013124003**

Contents

[Εύρεση και κατανόηση του χώρου εφαρμογής της ΑΔ. (5%) 2](#_Toc196148594)

[Εύρεση και κατανόηση των δεδομένων που θα αποθηκευτούν στην ΑΔ. (10%) 2](#_Toc196148595)

[Σχεδίαση της ΑΔ. (20%) 3](#_Toc196148596)

[Διαδικασία ETL (extract, transform, load) δηλ. «εξαγωγή, μετασχηματισμός και φόρτωση». Μαζί αυτές οι δραστηριότητες διαμορφώνουν τη διαδικασία που χρησιμοποιείται για τη λήψη δεδομένων από την πηγή και τη μετατροπή τους σε μία χρησιμοποιήσιμη μορφή και έπειτα μετακίνηση στην ΑΔ. (15%) 7](#_Toc196148597)

[Υλοποίηση της ΑΔ σε ένα από γνωστά συστήματα διαχείρισης βάσεων δεδομένων (SQL Server, MySQL, Postgres, Oracle). (20%) 4](#_Toc196148598)

[Διατύπωση (τουλάχιστον 6) ερωτημάτων SQL που θα χρησιμοποιούν συγκεντρωτικές συναρτήσεις και συνδυασμό πινάκων. (15%) 7](#_Toc196148599)

[Εκτέλεση των ερωτημάτων και εμφάνιση αποτελεσμάτων με οπτικοποίηση. (10%) 7](#_Toc196148600)

[Καταγραφή συμπερασμάτων. (5%) 8](#_Toc196148601)

# Εύρεση και κατανόηση του χώρου εφαρμογής της ΑΔ. (5%)

Σε αυτή την εργασία, θα ασχοληθούμε με δεδομένα οικονομικών συναλλαγών (Online, On site, etc)

# Εύρεση και κατανόηση των δεδομένων που θα αποθηκευτούν στην ΑΔ. (10%)

Τα δεδομένα βρίσκονται διαθέσιμα στο <https://www.kaggle.com/datasets/computingvictor/transactions-fraud-datasets/data> .  
Αυτό το σύνολο δεδομένων συνδυάζει αρχεία συναλλαγών, πληροφορίες πελατών και δεδομένα καρτών από ένα τραπεζικό ίδρυμα, που εκτείνονται σε όλη τη δεκαετία του 2010. Το σύνολο δεδομένων έχει σχεδιαστεί για πολλαπλούς αναλυτικούς σκοπούς, συμπεριλαμβανομένου του εντοπισμού απάτης, της ανάλυσης συμπεριφοράς πελατών και της πρόβλεψης εξόδων.  
  
Τα αποτελέσματα των ερωτημάτων και τα διαγράμματα βασίζονται στο αρχικό datataset των 1.3GB  
Στην εργασία, υπάρχει συννημένο έιναι μικρό υποσύνολο 60.000 εγγραγών , περίπου 6ΜΒ.  
  
2000 users.  
199 states.  
18638 cities.  
74831 merchants.  
197205 merchant locations.  
4136496 date records.  
4071 cards.  
3 chip usage types.  
13305915 transactions.

# Σχεδίαση της ΑΔ. (20%) Για την αποθήκη δεδομένων μας, θα χρησιμοποιήσουμε την [τοπολογία Αστέρα](https://en.wikipedia.org/wiki/Star_schema).

A screenshot of a computer

AI-generated content may be incorrect.

# Υλοποίηση της ΑΔ σε ένα από γνωστά συστήματα διαχείρισης βάσεων δεδομένων (SQL Server, MySQL, Postgres, Oracle). (20%)

Η βάση μας είναι υλοποιημένη σε Microsoft SQL Server 2022 και μπορεί να δημιουργηθεί εκτελώντας το αρχείο 01CreateDataWarehouse.sql

Για περισσότερες οδηγίες δείτε το README.MD

CREATE TABLE Dim\_User (

    UserId INT PRIMARY KEY,

    CurrentAge INT,

    RetirementAge INT,

    BirthYear INT,

    BirthMonth INT,

    Gender VARCHAR(10),

    Address VARCHAR(100),

    Latitude DECIMAL(9,6),

    Longitude DECIMAL(9,6),

    PerCapitaIncome DECIMAL(10,2),

    YearlyIncome DECIMAL(10,2),

    TotalDebt DECIMAL(10,2),

    CreditScore INT,

    NumberOfCreditCards INT

);

GO

CREATE TABLE Dim\_State (

    StateId INT IDENTITY(1,1) PRIMARY KEY,

    StateCode VARCHAR(50) UNIQUE

);

GO

CREATE TABLE Dim\_City (

    CityId INT IDENTITY(1,1) PRIMARY KEY,

    CityName VARCHAR(100),

    StateId INT,

    CONSTRAINT FK\_city\_state FOREIGN KEY (StateId) REFERENCES Dim\_State(StateId)

);

GO

CREATE TABLE Dim\_Merchant (

    MerchantId INT PRIMARY KEY,

    MCC INT

);

GO

CREATE TABLE Dim\_MerchantLocation (

    MerchantLocationId INT IDENTITY(1,1) PRIMARY KEY,

    MerchantId INT,

    CityId INT,

    CONSTRAINT FK\_merchant\_location\_merchant FOREIGN KEY (MerchantId) REFERENCES Dim\_Merchant(MerchantId),

    CONSTRAINT FK\_merchant\_location\_city FOREIGN KEY (CityId) REFERENCES Dim\_City(CityId),

    CONSTRAINT UQ\_merchant\_city UNIQUE (MerchantId, CityId)

);

GO

CREATE TABLE Dim\_Date (

    DateId DATETIME PRIMARY KEY,

    year INT,

    month INT,

    day INT,

    hour INT,

    minute INT

);

GO

CREATE TABLE Dim\_ChipUsage (

    ChipUsageId INT IDENTITY(1,1) PRIMARY KEY,

    UsageDescription VARCHAR(50) UNIQUE

);

GO

CREATE TABLE Dim\_Card (

    CardId INT PRIMARY KEY,

    UserId INT,

    CONSTRAINT FK\_dim\_card\_user FOREIGN KEY (UserId) REFERENCES Dim\_User(UserId)

);

GO

CREATE TABLE Fact\_Transaction (

    TransactionId INT PRIMARY KEY,

    DateId DATETIME,

    UserId INT,

    CardId INT,

    MerchantLocationId INT,

    amount DECIMAL(10,2),

    ChipUsageId INT,

    error VARCHAR(100),

    CONSTRAINT FK\_fact\_transaction\_date FOREIGN KEY (DateId) REFERENCES Dim\_Date(DateId),

    CONSTRAINT FK\_fact\_transaction\_user FOREIGN KEY (UserId) REFERENCES Dim\_User(UserId),

    CONSTRAINT FK\_fact\_transaction\_card FOREIGN KEY (CardId) REFERENCES Dim\_Card(CardId),

    CONSTRAINT FK\_fact\_transaction\_merchant\_location FOREIGN KEY (MerchantLocationId) REFERENCES Dim\_MerchantLocation(MerchantLocationId),

    CONSTRAINT FK\_fact\_transaction\_chip\_usage FOREIGN KEY (ChipUsageId) REFERENCES Dim\_ChipUsage(ChipUsageId)

);

GO

# Διαδικασία ETL (extract, transform, load) δηλ. «εξαγωγή, μετασχηματισμός και φόρτωση». Μαζί αυτές οι δραστηριότητες διαμορφώνουν τη διαδικασία που χρησιμοποιείται για τη λήψη δεδομένων από την πηγή και τη μετατροπή τους σε μία χρησιμοποιήσιμη μορφή και έπειτα μετακίνηση στην ΑΔ. (15%)

Έχουμε υλοποιήσει 2 stored procedures που διαβάζουν τα 2 .csv datasets (users, transactions) και εισάγουν τα δεδομένα στους πίνακες.  
Είναι αρκετά μεγάλες, εκτελέστε το αρχείο **02CreateStoredProcs.sql**

Μπορούμε να τις τρέξουμε manually **03ExecureStoredProcsToLoadData.sql**

EXEC sp\_load\_user\_data

    @dataFile = 'C:\dev\UTH\UTH-WAREHOUSE\users\_data.csv',

    @formatFile = 'C:\dev\UTH\UTH-WAREHOUSE\users\_format.fmt',

    @errorFile = 'C:\dev\UTH\UTH-WAREHOUSE\users\_errors.txt';

EXEC sp\_load\_transaction\_data

    @dataFile = 'C:\dev\UTH\UTH-WAREHOUSE\transactions\_data\_full.csv',

    @errorFile = 'C:\dev\UTH\UTH-WAREHOUSE\transactions\_errors.txt';

Είτε να φτιάξουμε ένα job το οποίο τρέχει κάθε βράδυ και εισάγει τα δεδομένα της ημέρες όπως στο **05CreateETLJob.sql**

# Διατύπωση (τουλάχιστον 6) ερωτημάτων SQL που θα χρησιμοποιούν συγκεντρωτικές συναρτήσεις και συνδυασμό πινάκων. (15%)

HJKHJKHJKHJKH

JKHJK

HJ

K

HJK

HJGK

HJ

K

HJK

HJK

# Εκτέλεση των ερωτημάτων και εμφάνιση αποτελεσμάτων με οπτικοποίηση. (10%)

HJKHJKHJKGHJK

HJK

HJKHJKH

# Καταγραφή συμπερασμάτων. (5%)

FGHFGH

FG

H

FGHFG