**Datawarehouse and Business Intelligence**

**IT20216900**

**De Silva S. R.**

# Table of contents

Contents

[Data Selection and introduction 2](#_Toc71651275)

[Preparation of the DataSources 3](#_Toc71651276)

[Class Diagrams Using source tables. 4](#_Toc71651277)

[Snowflake Schema 5](#_Toc71651278)

[Solution Architecture 6](#_Toc71651279)

Extraction from Source database to staging area. 8

[Data Profiling 9](#_Toc71651282)

[Transform and load data to tables. 10](#_Toc71651283)

[Slowly changing dimension 11](#_Toc71651284)

**Data Selection and Introduction**

The NFL Draft 1970 – 2021 dataset was chosen to complete the assignment. The original data has been modified and arranged in order to meet the given requirements. The data set contains the NFL Drafts since 1970. The dataset contains follows.

* Location –Contains Location of players and college. Location can have multiple players and colleges. This contains the hierarchy where one country has many states.
* College - Contains data of the college which includes capacity and the name of the college.
* Player - Contains data of players who plays pro-football such as player name, player position, age, and their manager.
* Draft – Contains the draft price and the goals scored by each player and the average goals scored within a month

Link to the dataset - [NFL Draft 1970-2021 | Kaggle](https://www.kaggle.com/datasets/cviaxmiwnptr/nfl-draft-19702021)

**Preparation of the Data Sources**

A database named NFL Draft 1979-2021 was created including the database, csv and txt source files.

* Dbo.Player
* Location.csv
* College.csv
* Draft.txt

A new database named NFL\_DW was built for the data warehouse which contains the dimensions and the fact table.

* DimLocation
* DimDate
* DimCollege
* DimPlayer
* FactDraft

The NFL\_Staging database was created in order to add the data to the database.

A script file was used to create the DimDate relation in the Data Warehouse.

**Class Diagram Using source tables**

**Shape

Description automatically generated with medium confidence**

**ER diagram**

Diagram

Description automatically generated

**Snowflake Schema**

For this assignment, I have used snowflake schema which aids in reducing redundancies through normalization. Location data is recorded in the locations table which is referred by the player and college tables.

Assumption - :

* PlayerPosition and Manager in player table are considered as Slowly changing dimensions

Diagram, schematic

Description automatically generated

**Solution Architecture**

Icon

Description automatically generated

Data Sources - : All the data

ETL -: ETL is process of extracting, transforming and loading.

Staging Area - : o

**Extraction from Source database to staging area**

Graphical user interface

Description automatically generated

**Data Profiling**

A screenshot of a computer

Description automatically generated with medium confidence

**Data warehouse design & development**

**A screenshot of a computer

Description automatically generated with medium confidence**

Stored Procedure for DimLocation

CREATE PROCEDURE dbo.UpdateDimLocation

@LocationID nvarchar(50),

@Country nvarchar(50),

@State nvarchar(50)

AS

BEGIN

if not exists (select LocationSK

from dbo.DimLocation

where AlternateLocationID = @LocationID)

BEGIN

insert into dbo.DimLocation

(AlternateLocationID, Country, State, InsertDate, ModifiedDate)

values

(@LocationID, @Country, @State, GETDATE(), GETDATE())

END;

if exists (select LocationSK

from dbo.DimLocation

where AlternateLocationID = @LocationID)

BEGIN

update dbo.DimLocation

set Country = @Country,

State = @State,

ModifiedDate = GETDATE()

where AlternateLocationID = @LocationID

END;

END;

Stored Procedure for DimCollege

CREATE PROCEDURE dbo.UpdateDimCollege

@CollegeID nvarchar(50),

@CollegeName nvarchar(50),

@Capacity int,

@LocationKey int

AS

BEGIN

if not exists (select CollegeSK

from dbo.DimCollege

where AlternateCollegeID = @CollegeID)

BEGIN

insert into dbo.DimCollege

(AlternateCollegeID, CollegeName, Capacity, LocationKey, InsertDate, ModifiedDate)

values

(@CollegeID, @CollegeName, @Capacity, @LocationKey, GETDATE(), GETDATE())

END;

if exists (select CollegeSK

from dbo.DimCollege

where AlternateCollegeID = @CollegeID)

BEGIN

update dbo.DimCollege

set CollegeName = @CollegeName,

Capacity = @Capacity,

LocationKey = @LocationKey,

ModifiedDate = GETDATE()

where AlternateCollegeID = @CollegeID

END;

END;

**Slowly Changing Dimension**

**Graphical user interface, application

Description automatically generated**

**ETL development – Accumulating fact tables**

A screenshot of a computer

Description automatically generated