Sri Lanka Institute of Information Technology

IT3021 – Data Warehousing and Business Intelligence

Year 3 – Semester 01

Individual Project

Assignment 1

Submitted by:

Name - Liyanage S.R

Registration Number - IT20005726

Batch Number – Y3.S1.WD.DS.05.1.G1

Step 1: Data set selection

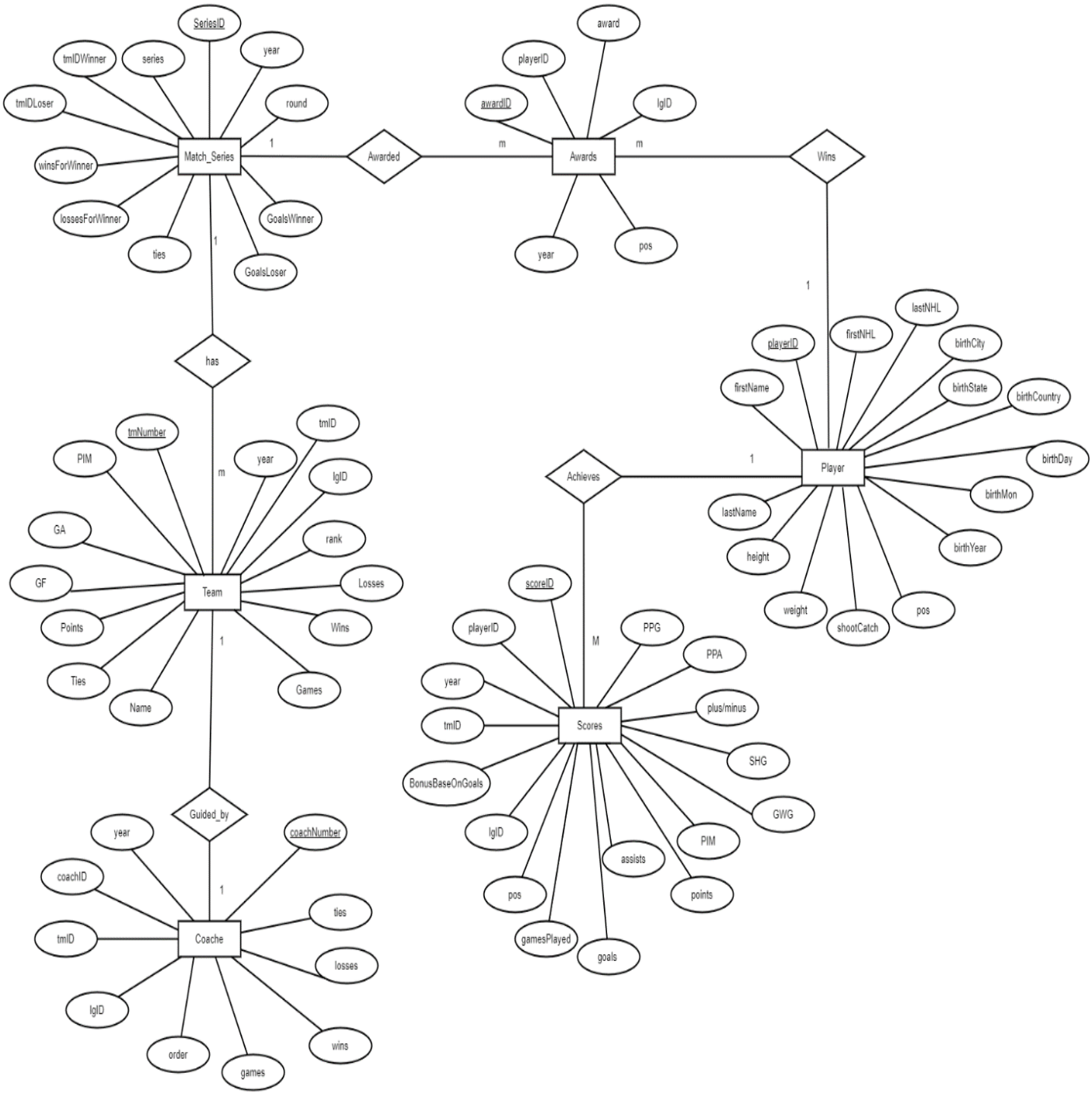
The data set I have chosen is about Professional Hockey Database. It contains data on hockey players, teams, and coaches from 1909 to 2011. The Hockey Database is a collection of statistics about professional hockey teams in North America Content. The data set had been chosen from Kaggle.com. The original dataset contains a huge amount of data files including players' biological information, scoring, awards, etc. However, I chose 6 files out of them namely master, AwardsPlayers, Scoring, Coaches, Teams, and SeriesPost. Those files also included a large amount of data, whereas some of those are unnecessary and some files contain a huge amount of null values. So I prepared the data set according to my requirements to obtain better hierarchies, dimensions, and aggregations.

Accordingly, the customized data source contains 6 tables. There is a players table that contains biological information of the players, an awards table that contains the awards won by players, Scoring table which contains all scoring statistics of the players. Furthermore, there is a match series table about the match series where the above awards are awarded to players, a teams' table which covers teams' details who are played in those match series. Finally, it contains a coaches table that describes the coaches who guided those teams.

Link for the original data set:

https://www.kaggle.com/datasets/open-source-sports/professional-hockey-database?select=SeriesPost.csv

ER Diagram:


SeriesID
year
series
tmIDWinner
tmIDLoser
round
Match_Series
1
Awarded
m
Awards
m
Wins
award
playerID
awardID
lgID
winsForWinner
lossesForWinner
GoalsWinner
ties
GoalsLoser
year
pos
1
1
has
lastNHL
playerID
firstNHL
birthCity
firstName
birthState
birthCountry
birthDay
1
Player
Achieves
birthMon
lastName
birthYear
height
weight
shootCatch
pos
tmNumber
tmID
PIM
year
lgID
m
GA
rank
GF
Losses
Team
Points
Wins
Ties
Games
Name
1
scoreID
playerID
PPG
PPA
M
year
plus/minus
tmID
Scores
SHG
BonusBaseOnGoals
GWG
lgID
PIM
assists
pos
points
gamesPlayed
goals
Guided_by
year
coachNumber
coachID
1
ties
tmID
Coache
losses
lgID
wins
order
games

Step 2: Preparation of data sources

The data set has been separated into multiple sources. There are 6 source files in three formats. The File formats which had been used are CSV, text files, and database table

Sources:

Database

Database Name - Professional_Hockey_SourceDB

Table name – Players

Players – Players' source describes the biological information of players including some details about their first and last hockey matches.

CSV files

Awards - Awards source describes the awards which had been awarded to the above hockey players.

Scoring – Scoring source describes the scores which had been achieved by those players in different hockey matches while playing in a team.

Text Files

Match Series – Match series source describes the details of the match series in which the above awards were awarded to players.

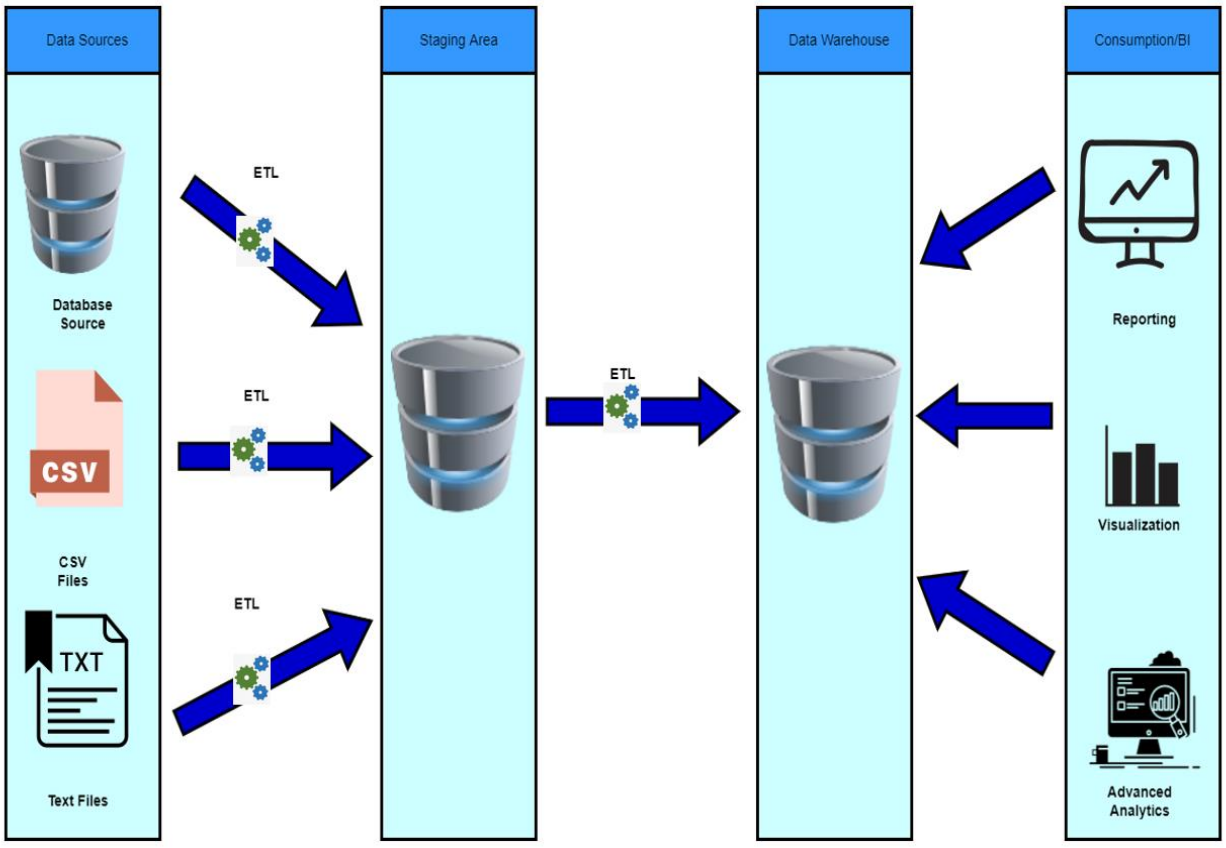
Teams – Teams source describes Teams' details who played in those match series.

Coaches – Coaches' source describes coaches who guided those teams.

Step 3: Solution architecture

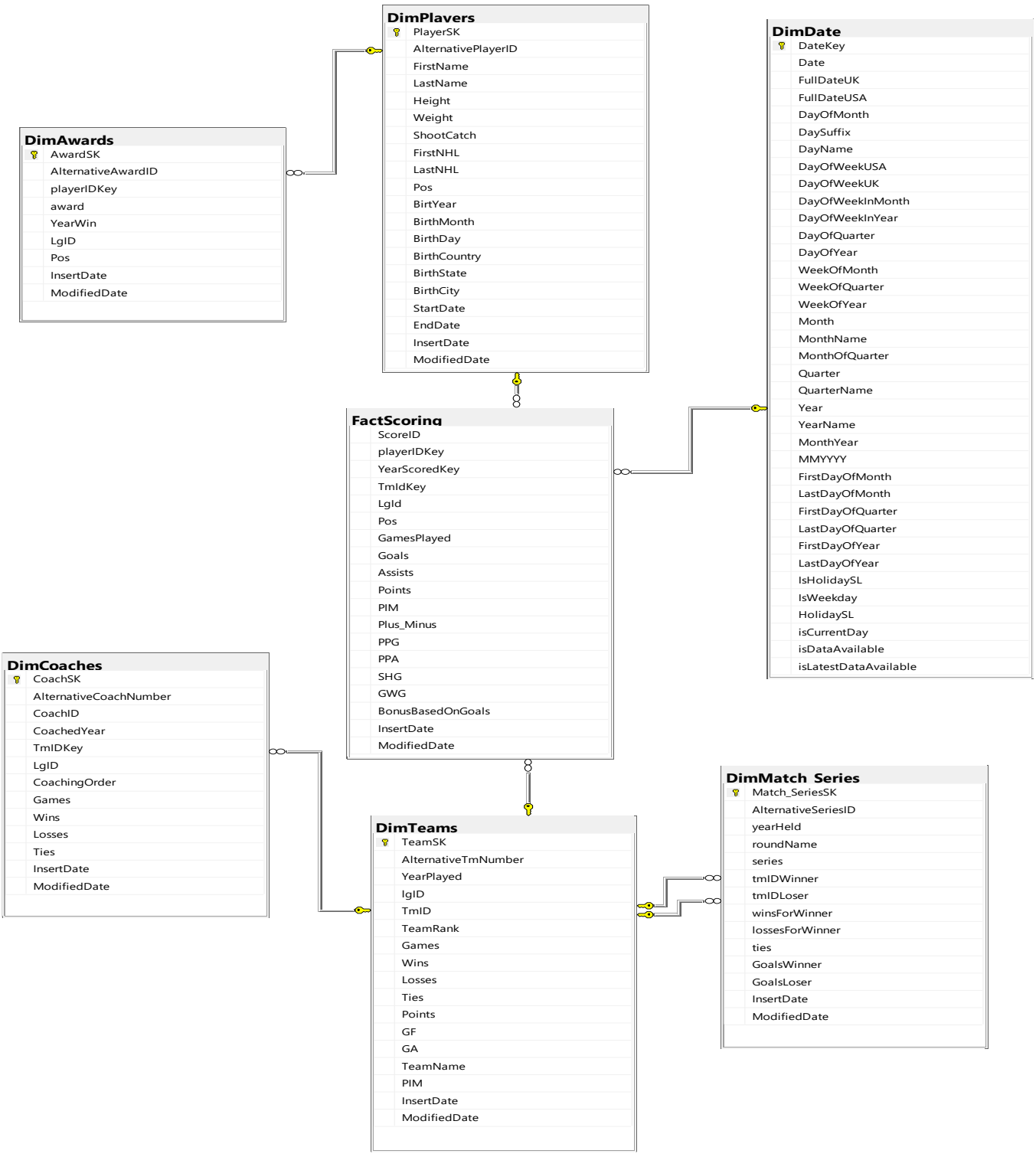
This is the Data Warehouse and BI solution architecture. There are four phases in this DW & BI solution architecture which lead to achieving better DW with a logical ETL process and finally a Better BI solutions.

- Data Sources - This describes the collection of data from various sources namely operational databases, Flat files (CSV and Text files). There are 6 data source tables included in this phase.
- Staging Area - Those data sources undergo processing through ETL processes which involved the Extraction of data from sources and getting into the Staging area.
- Data warehouse - After the successful staging phase, data should be loaded to Data Warehouse. Inhere, data that are in staging DB are extracted, transformed, and then Loaded to DW using available ETL methods while considering a logical order.
- Consumption/BI – In this phase, the data which were loaded to DW were used as the source for interpretations such as reporting, visualization, and advanced analytics.



Step 4: Data warehouse design & development

The Datawarehouse shema (dimensional model) chossen is snowflake.



There are six dimensions including the date dimension and one slowly changing dimension(DimPlayers). Also, there is one Fact table as well.

- DimPlayers: DimPlayers describes the biological information of players including some details about their first and last hockey matches. DimPlayers developed as a slowly changing dimension table. It has the playerSK as the primary key.

- DimTeams: This describes the teams which were played in the Match series. It has the teamSK as the primary key.

- DimMatchSeries: This describes the match series details. It has the Match_SeriesSK as the primary key. It takes teamSk as a foreign key to indicate winner teams and loser teams.

- DimCoaches: This describes the coaches who guided the teams. It has the coachSK as the primary key. It takes teamSK as a foreign key to indicate the team which was guided by the coach.

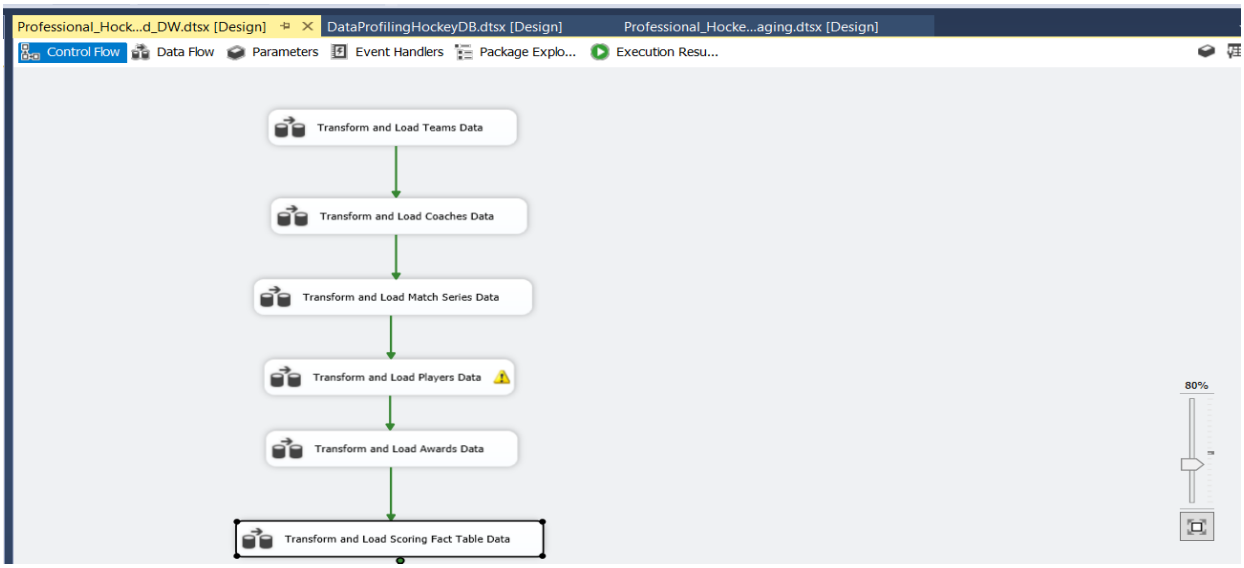
- DimAward: This describes the awards won by players. It has the awardSK as the primary key. It takes playerSK as a foreign key to indicate the player who wins that award.

- FactScoring: This describes the scores achieved by players. It takes playerSK as a foreign key to indicate which player achieved that score. Further, it takes DateKey as a foreign key to indicate the year when the match was held. It takes teamSk as a foreign key to indicate the team.

Step 5: ETL development

Extraction: In the SSIS, first of all, staging DB was created. Then data were extracted from all the above-mentioned sources and stored inside respective tables for each source.

Then, using staging DB as a source data was again extracted, transformed using necessary methods, and finally loaded to Data Warehouse using a logical order. All ETL processes are described below according to the order those data were loaded to the Data warehouse.

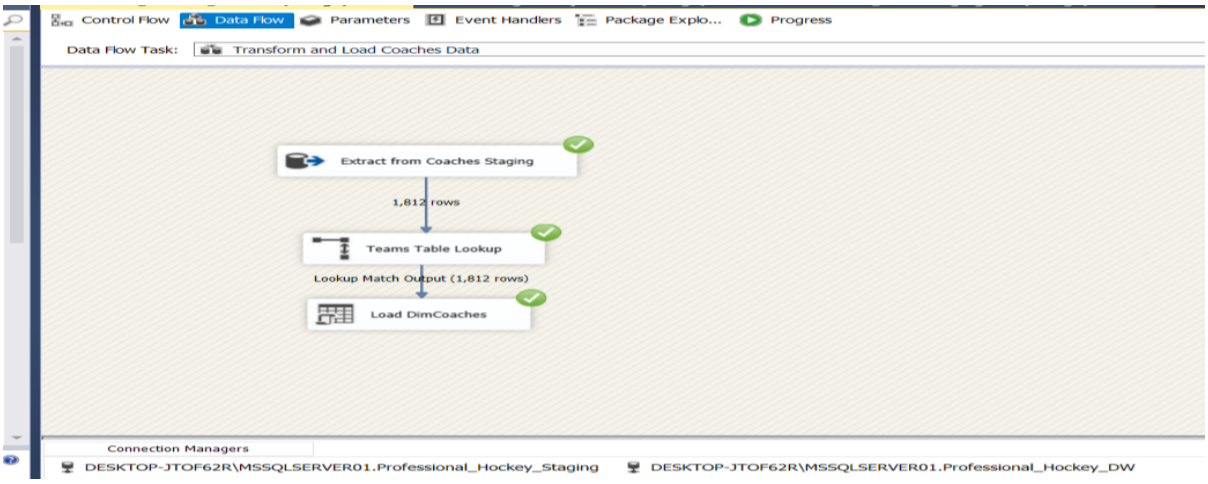


01. Extract,Transform and Load to DimTeams table

Extracted data from StgTeams table and Loaded to the DimTeams table of Data warehouse using a procedure(updateDimTeamsProcedure).

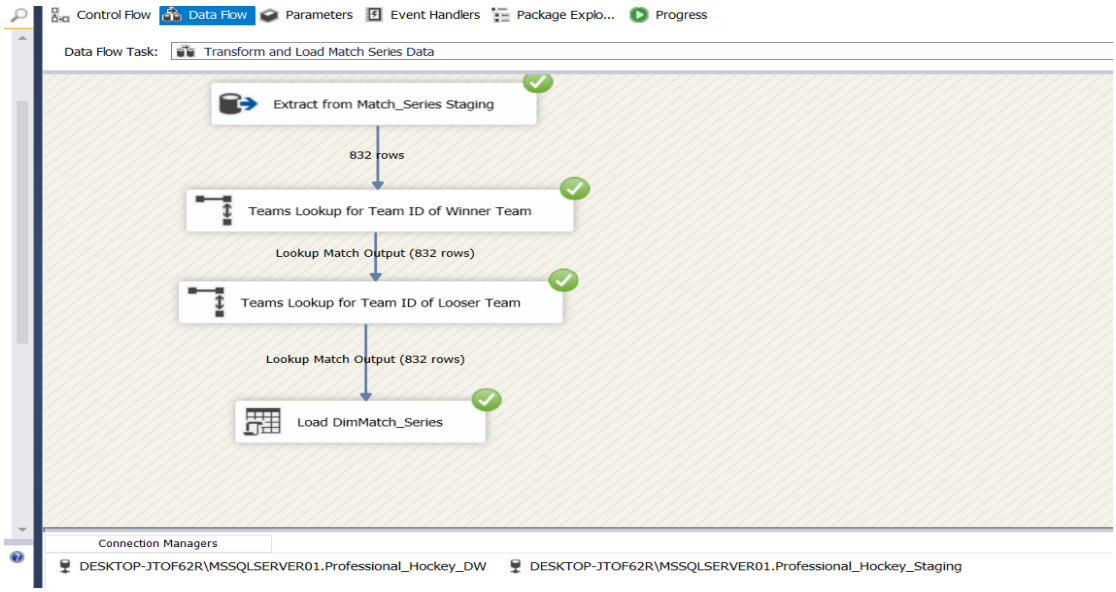
02. Extract,Transform and Load to DimCoaches table

Extracted data from StgCoaches table and use a lookup (Teams Lookup) to take the team table's surrogate key(teamSK). Then loaded to the DimCoaches table in Datawarehouse using a procedure(updateDimCoachesProcedure).



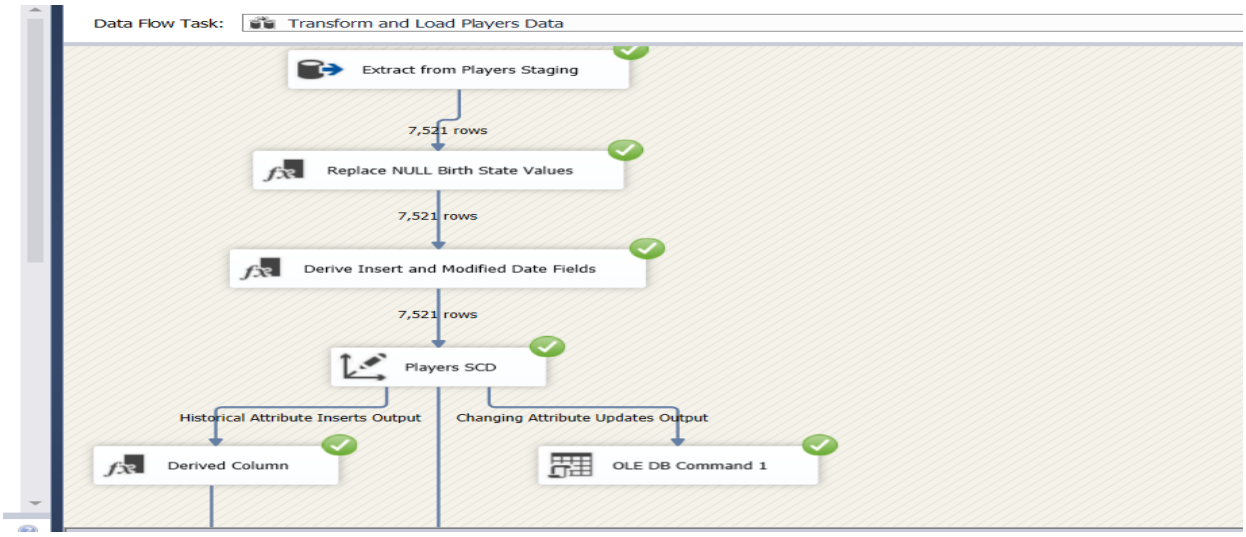
03. Extract,Transform and Load to DimMatch_Series table

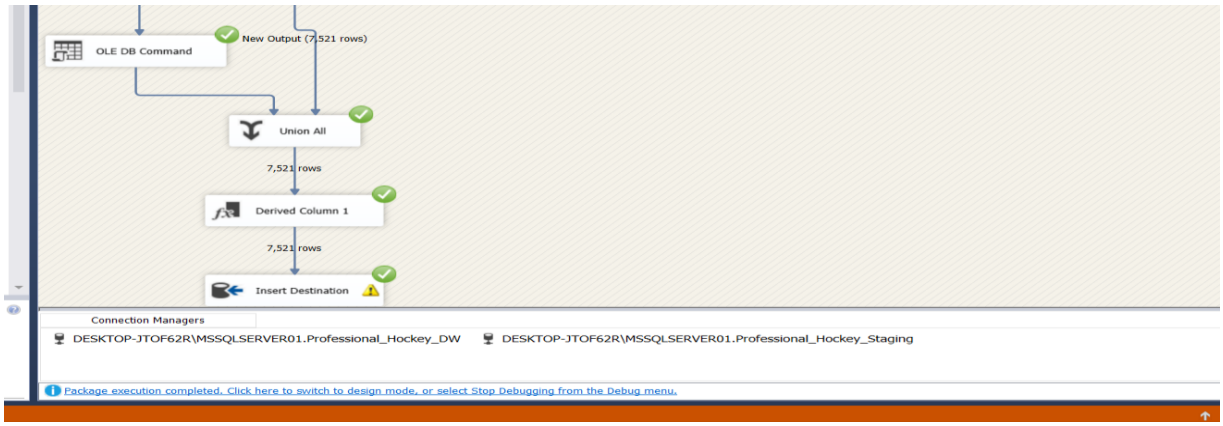
Extracted data from the StgMatch_Series table and use two lookups(Teams Lookup) for taking the teams table's surrogate key(teamSK) for the team id of the winner and team id of the looser. Then loaded to the DimMatch_Series table in Datawarehouse using a procedure (updateDimMatch_SeriesProcedure).



04. Extract, Transform and Load to DimPlayers table

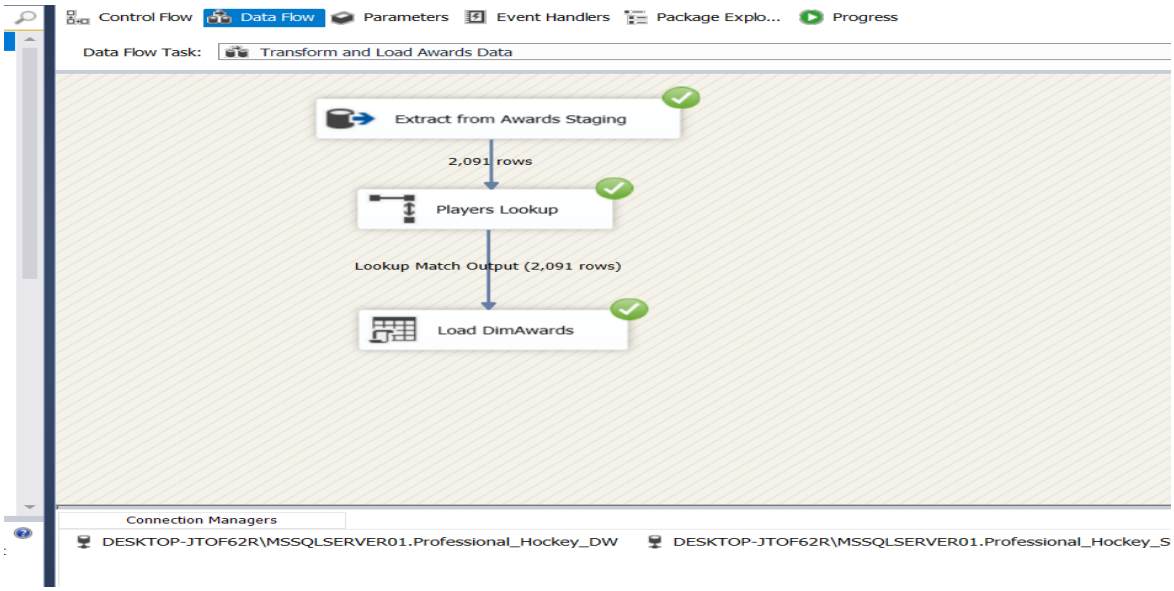
Extracted data from StgPlayers table and add a derived column to replace null values of the Birth country and Birth state columns. Then loaded players' data to DimPlayers table using a slowly changing dimension. In that case, height and weight were considered as changing attributes. Moreover, first NHL, last NHL, and last name took as historical attributes.





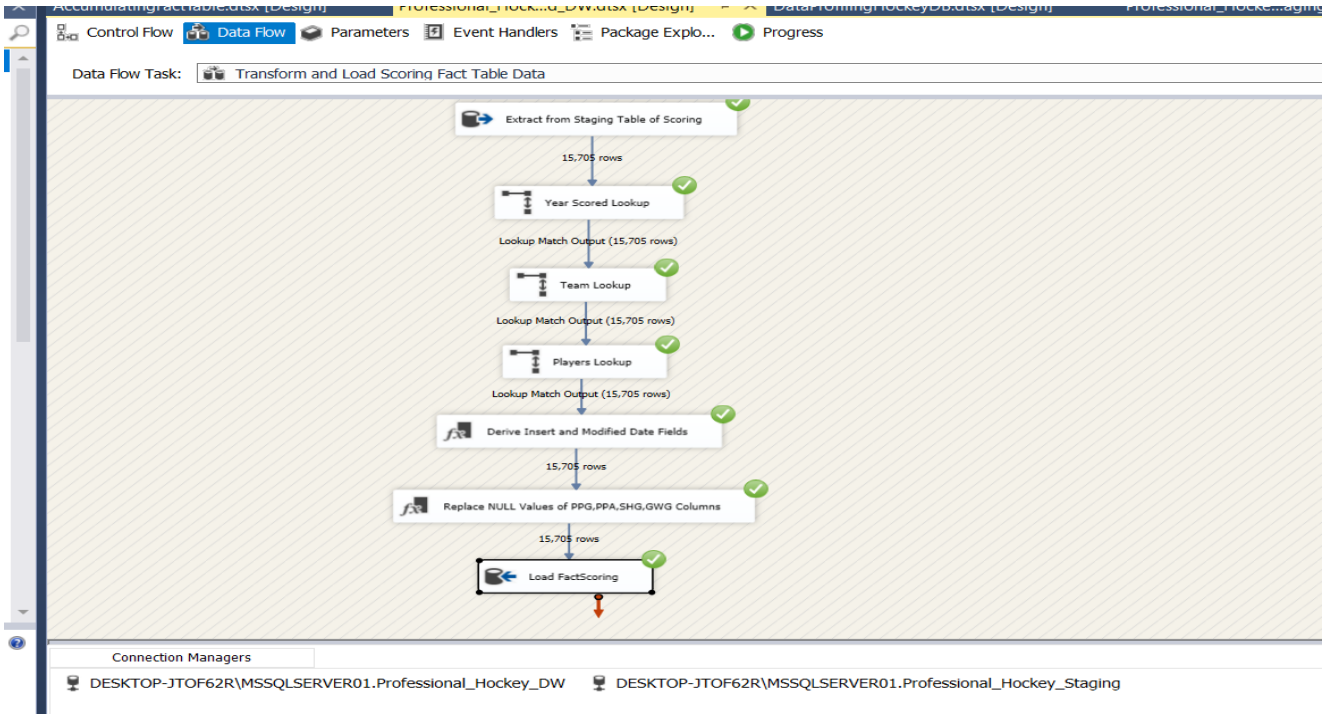
05. Extract, Transform and Load to DimAwards table

Extracted data from the StgAwards table and use a lookup(Players Lookup) for taking the players table's surrogate key(PlayerSK). Then loaded to the DimAwards table in Datawarehouse using a procedure (updateDimAwardsProcedure).



06. Extract,Transform and Load to FactScoring Table

Extracted data from the StgScoring table and added a lookup for taking the DimDate table's surrogate key to indicate the year scored. Then, another two lookups were added for taking the teams table's surrogate key(TeamSK) and the Players table's surrogate key to indicate the team ID and player ID respectively. Then a derived column is added to calculate the insert date and modified date. Thereafter, another derived column was added for replacing null values of a couple of columns namely PPG(power-play goals), PPA(power play assists), SHG(short-handed goals), and GWG(game-winning goals). Then loaded transformed data to the FactScoring table.


Control Flow
Data Flow
Parameters
Event Handlers
Package Explo...
Progress
Data Flow Task:
Transform and Load Scoring Fact Table Data
Extract from Staging Table of Scoring
15,705 rows
Year Scored Lookup
Lookup Match Output (15,705 rows)
Team Lookup
Lookup Match Output (15,705 rows)
Players Lookup
Lookup Match Output (15,705 rows)
Derive Insert and Modified Date Fields
15,705 rows
Replace NULL Values of PPG,PPA,SHG,GWG Columns
15,705 rows
Load FactScoring
Connection Managers
DESKTOP-JTOF62R\MSSQLSERVER01.Professional_Hockey_DW
DESKTOP-JTOF62R\MSSQLSERVER01.Professional_Hockey_Staging

Step 6: ETL development – Accumulating fact tables

In this step firstly new three columns were added to the fact table. Also, a new CSV file called Times was created to hold the ScoreId and accm_txn_complete_time.

Then another SSIS package was created for converting the fact table into an accumulating fact table. In that case, data that were in the Times CSV file was extracted using a flat-file source, and FactScoring tables' data was extracted using OLE DB Source. Then those sources were sorted by ScoreID and merged using a merge join. Then to count the complete time new derived column was added as 'adding process time' and the function was declared. Then data is loaded to the FactScoring table after mapping the correct parameters