****

**INTRODUCTION TO DATA MANAGEMENT PROJECT REPORT**

(Project Semester August-December 2020)

**OLYMPIC ATHLETE EVENTS**

Submitted by

J.SriramPrakash

Registration No:11809891

Programme and Section B.Tech CSE and KMO72

Course Code: INT 217

Under the Guidance of

**Vasudha ma’am with U.Id:23036**

**Discipline of CSE/IT**

**Lovely School of Computer science and engineering**

**Lovely Professional University, Phagwara**

**DECLARATION**

I, Sriramprakash, student of Introduction to data management under CSE/IT Discipline at, Lovely Professional University, Punjab, hereby declare that all the information furnished in this project report is based on my own intensive work and is genuine.

Date: 20-11-2020 Signature

Registration No. 11809891 J.sriramprakash

**Acknowledgement**

The success and final outcome of this project required a lot of guidance and assistance from my faculty. All I have done is only because such supervision and assistance and I would not forget to thank her.

I respect and thank my faculty, for providing me an opportunity to do the **Data management Project** and giving me all support and guidance which helped me to complete the project duty. I am extremely thankful to her for providing such a good support and guidance.

**Sriramprakash**

TABLE OF CONTENTS:

[**SECTON1: INTRODUCTION** 5](#_Toc57565733)

[**SECTION2: Objectives/Scope of the Analysis:** 7](#_Toc57565734)

[**SECTION3: Source of dataset:** 7](#_Toc57565735)

[**SECTION4: ETL process: -** 7](#_Toc57565736)

[What is ETL? 8](#_Toc57565737)

[**1) Extraction** 8](#_Toc57565738)

[**2) Transformation** 9](#_Toc57565739)

[**3) Loading** 9](#_Toc57565740)

[ETL in Data Warehouses 10](#_Toc57565741)

[**ETL Tools for Data Warehouses** 10](#_Toc57565742)

[**Who uses ETL?** 10](#_Toc57565743)

[**Why use ETL?** 10](#_Toc57565744)

[**Benefits of ETL** 11](#_Toc57565745)

[**SECTION5: Analysis on dataset:** 11](#_Toc57565746)

[Introduction: 11](#_Toc57565747)

[General Description: 12](#_Toc57565748)

[Specific Requirements, functions and formulas: 12](#_Toc57565749)

[**SECTION6: Analysis results: -** 13](#_Toc57565750)

[**SECTION7: Visualization: -** 13](#_Toc57565751)

[**SECTION8:References:** 21](#_Toc57565752)

# **SECTON1: INTRODUCTION**

I have undertaken a project titled “Olympic Athlete Events”.

The first modern Olympics were held in Athens, Greece, in 1896. In the opening ceremony, King Georgios I and a crowd of 60,000 spectators welcomed 280 participants from 13 nations (all male), who would compete in 43 events, including track and field, gymnastics, swimming, wrestling, cycling, tennis, weightlifting, shooting and fencing. All subsequent Olympiads have been numbered even when no Games take place (as in 1916, during [World War I](https://www.history.com/topics/world-war-i), and in 1940 and 1944, during [World War II](https://www.history.com/topics/world-war-ii)). The official symbol of the modern Games is five interlocking colored rings, representing the continents of North and South America, Asia, Africa, Europe and Australia. The Olympic flag, featuring this symbol on a white background, flew for the first time at the Antwerp Games in 1920.

The Olympics truly took off as an international sporting event after 1924, when the VIII Games were held in Paris. Some 3,000 athletes (with more than 100 women among them) from 44 nations competed that year, and for the first time the Games featured a closing ceremony. The Winter Olympics debuted that year, including such events as figure skating, ice hockey, bobsledding and the biathlon. Eighty years later, when the 2004 Summer Olympics returned to Athens for the first time in more than a century, nearly 11,000 athletes from a record 201 countries competed. In a gesture that joined both ancient and modern Olympic traditions, the shotput competition that year was held at the site of the classical Games in Olympia.

The first written records of the ancient Olympic Games date to 776 B.C., when a cook named Coroebus won the only event–a 192-meter footrace called the stade (the origin of the modern “stadium”)–to become the first Olympic champion. However, it is generally believed that the Games had been going on for many years by that time. Legend has it that Heracles (the Roman [Hercules](https://www.history.com/topics/ancient-history/hercules)), son of Zeus and the mortal woman Alcmene, founded the Games, which by the end of the 6th century B.C had become the most famous of all Greek sporting festivals. The ancient Olympics were held every four years between August 6 and September 19 during a religious festival honoring Zeus. The Games were named for their location at Olympia, a sacred site located near the western coast of the Peloponnese peninsula in southern Greece. Their influence was so great that ancient historians began to measure time by the four-year increments in between Olympic Games, which were known as Olympiads.

After 13 Olympiads, two more races joined the stade as Olympic events: the diaulos (roughly equal to today’s 400-meter race), and the dolichos (a longer-distance race, possibly comparable to the 1,500-meter or 5,000-meter event). The pentathlon (consisting of five events: a foot race, a long jump, discus and javelin throws and a wrestling match) was introduced in 708 B.C., boxing in 688 B.C. and chariot racing in 680 B.C. In 648 B.C., pankration, a combination of boxing and wrestling with virtually no rules, debuted as an Olympic event. Participation in the ancient Olympic Games was initially limited to freeborn male citizens of Greece; there were no women’s events, and married women were prohibited from attending the competition.

The seasons: -

The Winter Olympics occurs during the season when all the sports which require ice and snow take place. The Summer Olympics occur in summer, which enables more sporting events due to the conducive weather conditions. With more sporting events, the Summer Olympics also has more fans watching worldwide, as compared to the Winter Olympics.

The Summer Olympics is a more significant event, as it also holds a broader variety of five categories of sporting events. These include athletics, badminton, basketball, rugby, boxing, cricket, cycling, swimming, gymnastics, football, table tennis, hockey, handball, wrestling, and water motorsports, among many others.

The Winter Olympics hosts fewer sports broken down into three main groups. These events include ice hockey, snowboarding, figure skating, bobsleigh and ski jumping, among others. The Summer Olympics hosts 300 events covering 28 sports, while the Winter Olympics has 102 sporting events over 15 sports. By having more sporting events, as well as competitions, the Summer Olympics has a more significant number of athletes participating.

This dataset provides athlete and event data for all Olympics held from 1896 to 2016, the Winter and Summer Games were held in the same year up until 1992. After that, they began to be staggered, such that the Summer and Winter games alternate on a four-year cycle

# 

# **SECTION2: Objectives/Scope of the Analysis:**

The main objectives of my project are to know:

* + - 1. The most participated countries.
      2. Number of medals won by a country.
      3. Percentage of males and females participated in the Olympic over the years.
      4. Most medals won by a team.
      5. Participants avg height and weight.
      6. Top 20 events based on their occurrence.
      7. No. of times a team participated.

# **SECTION3: Source of dataset:**

[**https://knowledge.domo.com/Training/Self-Service\_Training/Onboarding\_Resources/Fun\_Sample\_Datasets**](https://knowledge.domo.com/Training/Self-Service_Training/Onboarding_Resources/Fun_Sample_Datasets)

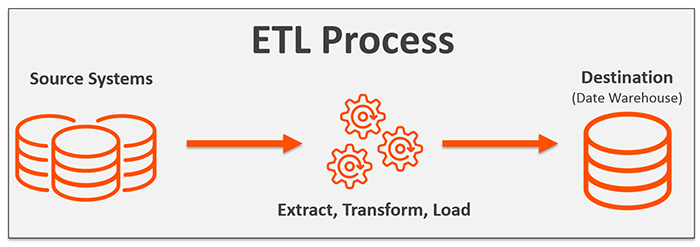
# **SECTION4: ETL process: -**

## What is ETL?

The process of extracting data from multiple source systems, transforming it to suit business needs, and loading it into a destination database is commonly called ETL, which stands for extraction, transformation, and loading. While ETL is usually explained as three distinct steps, this actually simplifies it too much as it is truly a broad process that requires a variety of actions.

ETL first saw a rise in popularity during the 1970s, when organizations began to use multiple databases to store their information. It quickly became the standard method for taking data from separate sources, transforming it, and loading it to a destination. A few decades later, data warehouses became the next big thing, providing a distinct database that integrated information from multiple systems. In order to accommodate our ever-changing world of digital technology in recent years, the number of data systems, sources, and formats has exponentially increased, but the need for ETL has remained just as important for an organization’s broader data integration strategy.

The following tasks are the main actions that happen in the ETL process:



### **1) Extraction**

The first step in ETL is extraction. During extraction, data is specifically identified and then taken from many different locations, referred to as the Source. The Source can be a variety of things, such as files, spreadsheets, database tables, a pipe, etc. It is not typically possible to pinpoint the exact subset of interest, so more data than necessary is extracted to ensure it covers everything needed. The volume of data extracted greatly varies and depends on business needs and requirements. Some extractions consist of hundreds of kilobytes all the way up to gigabytes. This is also the case for the timespan between two extractions; some may vary between days or hours to almost real-time.

Data extraction most typically occurs in one of three ways:

1. Update notification – the system notifies you when a record has been changed. This is typically referred to as the easiest method of extraction.
2. Incremental extraction – some systems cannot provide notifications for updates, so they identify when records have been modified and provide an extract on those specific records
3. Full extraction – some systems aren’t able to identify when data has been changed at all, so the only way to get it out of the system is to reload it all. This is usually only recommended for small amounts of data as a last resort

### **2) Transformation**

The next step in the ETL process is transformation. After data is extracted, it must be physically transported to the target destination and converted into the appropriate format. This data transformation may include operations such as cleaning, joining, and validating data or generating calculated data based on existing values. Involves converting the data from legacy format to warehouse format.

Whether the transformation takes place in the data warehouse or beforehand, there are both common and advanced transformation types that prepare data for analysis. Some of these include:

* Basic transformations:
  + Cleaning
  + Format revision
  + Restructuring
  + Deduplication
* Advanced transformations:
  + Filtering
  + Joining
  + Splitting
  + Derivation
  + Summarization
  + Integration

### **3) Loading**

* The final step in the ETL process involves loading the transformed data into the destination target. This target may be a database or a data warehouse. There are two primary methods for loading data into a warehouse: full load and incremental load. The full load method involves an entire data dump that occurs the first time the source is loaded into the warehouse. The incremental load, on the other hand, takes place at regular intervals. These intervals can be streaming increments (better for smaller data volumes) or batch increments (better for larger data volumes).
* Involves sorting, summarizing, consolidating, checking integrity, and building indices and partitions

## ETL in Data Warehouses

Data warehouse, A data warehouse is constructed by integrating data from multiple heterogeneous sources. It supports analytical reporting, structured and/or ad hoc queries and decision making. This tutorial adopts a step-by-step approach to explain all the necessary concepts of data warehousing. A data warehouse essentially combines information from several sources into one comprehensive database. For example, in the business world, a data warehouse might incorporate customer information from a company's point-of-sale systems (the cash registers), its website, its mailing lists and its comment cards

For a majority of companies, it is extremely likely that they will have years and years of data and information that needs to be stored. In order to consolidate all of this historical data, they will typically set up a data warehouse where all of their separate systems end up. Combining all of this information into one place allows easy reporting, planning, data mining, etc. Due to the fact that all of the data sources are different, as well as the specific format that the data is in may vary, their next step is to organize an ETL system that helps convert and manage the data flow.

In order to keep everything up-to-date for accurate business analysis, it is important that you load your data warehouse regularly. This means that all operational systems need to be extracted and copied into the data warehouse where they can be integrated, rearranged, and consolidated, creating a new type of unified information base for reports and reviews.

**ETL Tools for Data Warehouses**

While you can design and maintain your own ETL process, it is usually considered one of the most challenging and resource-intensive parts of the data warehouse project, requiring a lot of time and labor. Many organizations utilize ETL tools that assist with the process, providing capabilities and advantages unavailable if you were to complete it on your own. These tools can not only support with the extraction, transformation and loading process, but they can also help in designing the data warehouse and managing the data flow.

**Who uses ETL?**

ETL tools are often visual design tools that allow companies to build the program visually, versus just with programming techniques. For the most part, enterprises and companies that need to build and maintain complex data warehouses will invest in ETL and ETL tools, but other organizations may utilize them on a smaller scale, as well.

**Why use ETL?**

Since it was first introduced almost 50 years ago, businesses have relied on the ETL process to get a consolidated view of their data. ETL allows organizations to analyze data that resides in multiple locations in a variety of formats, streamlining the reviewing process and driving better business decisions.

**Benefits of ETL**

* Transforms data from multiple sources and loads it into various targets
* Provides deep historical context for businesses
* Allows organizations to analyze and report on data more efficiently and easily
* Increases productivity as it quickly moves data without requiring the technical skills of having to code it first
* Evolves and adapts to changing technology and integration guidelines

# **SECTION5: Analysis on dataset:**

## Introduction:

In this project I used various topics that I have learnt from this course. This projectt named “OLYMPIC ATHLETE EVENTS” is about the events occurred and medals own by each time or nation in respected sport or events took place over the years.

The topics I used in this project are ETL, Hyperlinking, Pivoting and Dashboard. The data I got from the online websites may contain the various mistakes such as data type of a columns or spelling mistakes, so to resolve the issue I used the ETL process. After the ETL process the dataset of 15 columns was reduced to 14 columns.

## General Description:

After the ETL process The dataset reduced to 14 columns namely: ID, Name, Year, Sex, Age, Country, Team, Height, Weight, Games, Season, City, Sport, Event, Medal.

ID: - An ID number assigned to this athlete based on their sequential order in the dataset.

Name: - The name of the athlete.

Year: - The year of the event.

Sex: - The gender of the athlete.

Age: - The age of the athlete.

Country: - The three-letter abbreviation for the country represented by the athlete.

Team: - The country this athlete represents.

Height: - The height of the athlete, in centimeters.

Weight: - The weight of the athlete, in kilograms.

Games: - The year and season for this Olympic event.

Season: - The season of the event (either Summer or Winter).

City: -The city in which these Olympics was held.

Sport: -The sport of the event.

Event: -The name of the event.

Medal: -The medal won by the athlete ("NA" if no medal was won).

## Specific Requirements, functions and formulas:

To do the cleaning/ETL process I required a tool tableau prep in my project. The functions and the thee formulas I used in this project are Pivoting, Hyperlinking, and the dashboard.

**Graphical user interface, application

Description automatically generated**

# **SECTION6: Analysis results: -**

1. The most participated country: -USA
2. Most Number of medals won by a country: - USA
3. Percentage of males and females participated in the Olympic over the years: - 73%of men and 23% of women
4. Most medals won by a team: - USA
5. Participants avg height and weight: - 175 cm in height and 71 kgs in weight.
6. Top 20 events based on their occurrence.
7. No. of times a team participated.

# **SECTION7: Visualization: -**

**Graphical user interface, application, PowerPoint

Description automatically generated**

**Fig-1.1)DASHBOARD**

Graphical user interface, application, table, Excel

Description automatically generated **Fig:1.2)HYPERLINKING**

A screenshot of a computer

Description automatically generated

**Fig:1.3) MOST PARTICIPATED COUTNRIES.**

Graphical user interface, application, table

Description automatically generated

**Fig:1.4)NUMBER OF MEDALS WON BY A COUNTRY.**

**Graphical user interface, application, table, Excel, PowerPoint

Description automatically generated**

**Fig:1.5)PERCENTAGE OF MALE AND FEMALE.**

Graphical user interface, application

Description automatically generated

**Fig-1.6)PARTICIPANTS AVG HEIGHT AND WEIGHT.**

**Graphical user interface, application, table

Description automatically generated**

**Fig:1.7)TOP 20 EVENTS.**

**Graphical user interface, application

Description automatically generated**

**Fig:1.8)NO OF TIMES A TEAM PARTICIPATED.**

# **SECTION8:References:**

[**https://knowledge.domo.com/Training/Self-Service\_Training/Onboarding\_Resources/Fun\_Sample\_Datasets**](https://knowledge.domo.com/Training/Self-Service_Training/Onboarding_Resources/Fun_Sample_Datasets)

# Works Cited

**lecturer:-Vasudha, source:- https://knowledge.domo.com/Training/Self-Service\_Training/Onboarding\_Resources/Fun\_Sample\_Datasets, references: Google, PPTs, samples datasets done in the class.**