**Data Project Proposal: Olympic Athlete Performance Analysis**

**Step 1: Preparing for Your Proposal**

**Client and Dataset Selection**

**Client:** The selected dataset selected by me is athlete\_events.csv, **National Olympic Committee (NOC)** or a **Sports Analytics Organization**. These entities are keenly interested in understanding historical performance trends, athlete demographics, and strategic insights to optimize future training programs, resource allocation, and talent identification.

**Dataset:** The selected dataset is athlete\_events.csv, which contains comprehensive historical data from the Olympic Games. This dataset is ideal because it offers a rich collection of information on individual athletes, their participation in various events, team affiliations, and medal outcomes across multiple Olympic cycles. This wealth of detail allows for a deep dive into performance metrics, demographic shifts, and country-specific achievements over time.

**Data Import and Cleaning Steps (Conceptual for SSMS)**

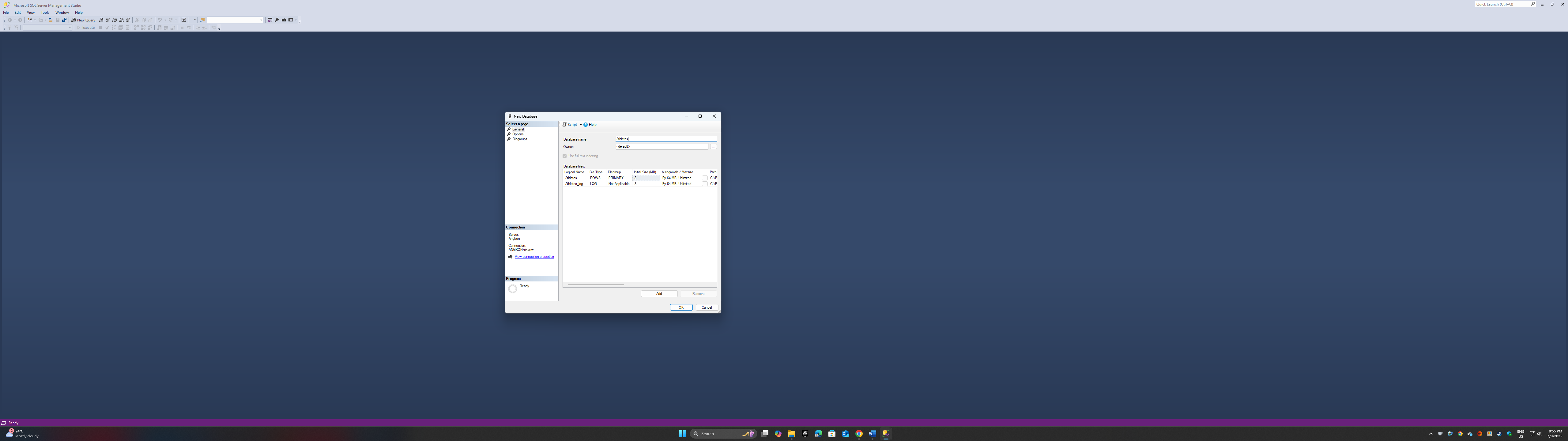
To prepare the athlete\_events.csv data for analysis in SQL Server Management Studio (SSMS), the following steps would typically be undertaken: 

Fig 1:- A Database is created in SSMS

1. **Import Data:**
   * Open SSMS and connect to the desired SQL Server instance.
   * Converted the CSV file to Excel file by saving as in excel
   * Right-click on the target database, navigate to "Tasks" -> "Import Excel file..." or "Import Data...".
   * Select athlete\_events.xlsx as the source file.
   * Review the proposed table name (e.g., athlete\_events) and schema. Ensured data types are correctly inferred (e.g., INT for ID, VARCHAR for Name, DECIMAL for Height/Weight, INT for Year).
   * Completed the import wizard.

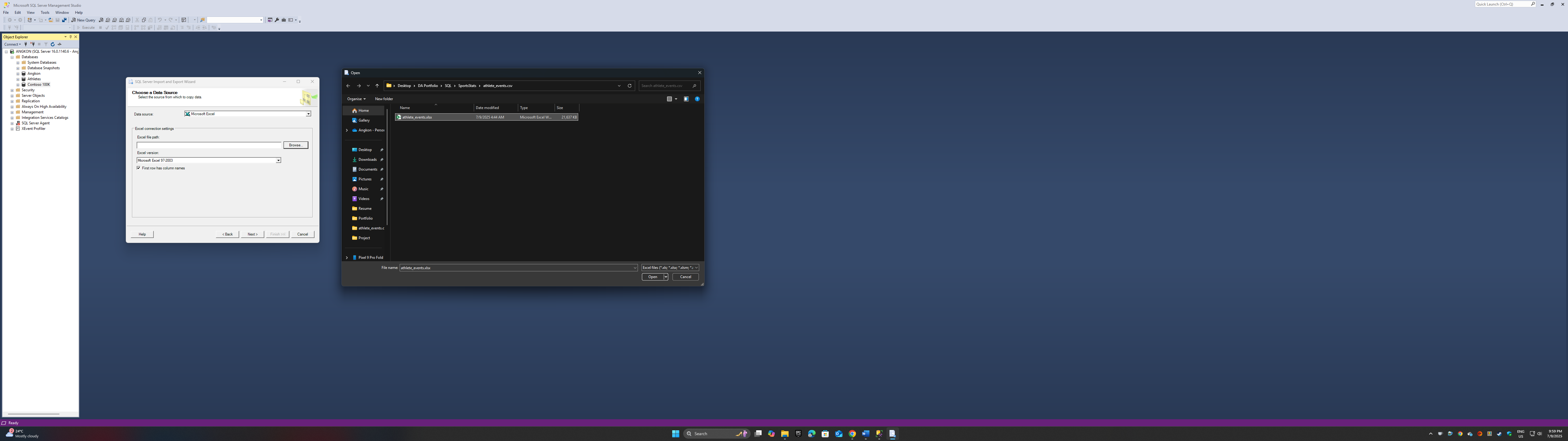


Fig 2:- Importing data in the database created earlier in SSMS

1. **Initial Data Cleaning and Transformation:**
   * **Handle Missing Values:**
     + For numerical columns like Age, Height, and Weight, identify rows where these values are NULL. Depending on the analysis, these might be excluded, imputed (e.g., with averages for specific sports/sexes), or handled during queries (e.g., WHERE Age IS NOT NULL).
     + For Medal, NULL indicates no medal was won. This is a valid state and should be preserved.
   * **Consistency Checks:**
     + Examine Team and NOC columns for potential inconsistencies or variations that might require standardization.
     + Check for duplicate ID and Name combinations, though an athlete might have multiple entries for different games/events. The ID column is likely a unique athlete identifier.
   * **Primary Key Consideration:** While the ID column serves as an athlete identifier, for a single denormalized table, a composite primary key might be considered if each row truly represents a unique participation instance (e.g., ID, Games, Event). For a normalized schema (as proposed in the ERD), individual primary keys will be assigned.

**Initial Data Exploration (SQL Queries for SSMS)**

A computer screen with a white screen

AI-generated content may be incorrect.

**---- View the first 20 rows to understand the table structure and data types**

SELECT TOP 20 \*

FROM athlete\_events

**-- Get the total number of records =271116**

SELECT COUNT(\*) AS total\_records

FROM athlete\_events

**-- Count Total athletes=135571**

SELECT COUNT(DISTINCT ID) AS Total\_Athletes

FROM athlete\_events

**-- Count total unique sports=66**

SELECT COUNT(DISTINCT Sport) AS Total\_Sport

FROM athlete\_events

**-- List all unique sports**

SELECT DISTINCT Sport

FROM athlete\_events

**-- List all unique season=2**

SELECT DISTINCT Season

FROM athlete\_events

ORDER BY Season;

**--Range of Years Covered =2016/1896**

SELECT MAX(Year) AS Latest\_Year,

MIN(Year) AS Last\_Year

FROM athlete\_events

**-- Count the distribution of medals = NA-231333,Gold-13372,Bronze-13295,Silver-13116**

SELECT Medal,

COUNT(\*) AS Medal\_count

FROM athlete\_events

GROUP BY Medal

ORDER BY Medal\_count DESC

**-- Calculate average age, height, and weight of participants-Age 25.055,Height-175.371,Weight-70.688**

SELECT

AVG(Age) AS Average\_Age,

AVG(Height) AS Average\_Height,

AVG(Weight) AS Average\_Weight

FROM athlete\_events

WHERE Age IS NOT NULL AND Height IS NOT NULL AND Weight IS NOT NULL

**-- Top 10 National Olympic Committees (NOCs) by total medal count=** 1.USA,2.FRA,3.GBR....

SELECT TOP 10 NOC,

COUNT(\*) AS Best

FROM athlete\_events

WHERE Medal IS NOT NULL

GROUP BY NOC

ORDER BY Best DESC

**Proposed Entity Relationship Diagram (ERD)**

The athlete\_events.csv The dataset is initially a single, denormalized table. For a more robust and scalable data analysis project, we propose normalizing the data into several interconnected tables. This approach reduces data redundancy, improves data integrity, and facilitates more complex and efficient querying. A diagram of a data flow

AI-generated content may be incorrect.

Relationships:

- Athletes (1) -- (\*M\*) Participations: An athlete can have many participations.

- Games (1) -- (\*M\*) Participations: A game can have many participations.

- Events (1) -- (\*M\*) Participations: An event can have many participations.

- Teams (1) -- (\*M\*) Participations: A team can have many participations.

- Sports (1) -- (\*M\*) Events: A sport can have many events.

**Description of Tables:**

* **Athletes:** Stores unique information about each athlete (e.g., ID, Name, Sex, Height, Weight). AthleteID is the primary key.
* **Games:** Stores details about each Olympic Games instance (e.g., Year, Season, Host City). GameID is the primary key.
* **Sports:** Stores unique names of sports (e.g., Athletics, Swimming). SportID is the primary key.
* **Events:** Stores unique names of events within sports (e.g., 100m Dash, Marathon). It has a foreign key SportID linking it to the Sports table. EventID is the primary key.
* **Teams:** Stores information about the teams/NOCs (e.g., Team Name, NOC abbreviation). TeamID is the primary key.
* **Participations (Fact Table):** This is the central fact table that records each instance of an athlete participating in a specific event during a particular game. It contains foreign keys linking to Athletes, Games, Events, and Teams, along with specific details for that participation like Age (as age can change across games) and Medal outcome. ParticipationID is the primary key.

This normalized schema provides a robust foundation for complex queries and ensures data integrity.

**Step 2: Develop Project Proposal**

**Description**

This project aims to provide actionable insights into historical Olympic athlete performance and trends, leveraging a comprehensive dataset of past Olympic Games. The analysis will focus on identifying patterns in medal distribution, exploring the evolution of athlete demographics (age, height, weight) over time, and assessing the performance of various National Olympic Committees (NOCs) across different sports and seasons. The findings will be of significant interest to the National Olympic Committees seeking to optimize athlete development strategies, sports federations aiming to understand competitive landscapes, and sports scientists studying long-term athletic trends. Our primary audience will be decision-makers within these sports organizations, including coaches, strategists, and policymakers.

**Questions**

1. **Which National Olympic Committees (NOCs) have demonstrated the most consistent and dominant performance in specific sports or overall, across different Olympic eras (e.g., pre-1980 vs. post-1980)?** This question will help identify historical powerhouses and emerging nations in the Olympic landscape.
2. **How have the average physical characteristics (age, height, weight) of medal-winning athletes changed over time within specific sports (e.g., gymnastics vs. basketball), and does this correlate with performance improvements?** This will explore the evolution of the ideal athlete profile in various disciplines.
3. **What is the distribution of medals across different sports and events for both Summer and Winter Olympics, and are there any sports that consistently contribute a disproportionate number of medals to certain NOCs?** This question will highlight the strategic importance of certain sports for national medal counts.

**Hypothesis**

1. **Hypothesis 1:** Countries with larger populations and higher GDPs will generally accumulate more Olympic medals overall, but smaller, specialized nations may dominate specific niche sports due to focused investment and talent development.
2. **Hypothesis 2:** Over time, the average age of Olympic medalists in physically demanding sports (e.g., swimming, gymnastics) has decreased, while in technically demanding or endurance sports (e.g., shooting, marathon), it may have increased, reflecting specialized training and career longevity.
3. **Hypothesis 3:** There will be a clear shift in the distribution of medals from traditional Western powers to emerging Asian and African nations in certain sports over the last few decades, indicating a globalization of athletic excellence.

**Approach**

Our approach will strictly utilize SQL queries within SSMS to analyze the athlete\_events dataset, potentially after normalizing it into the proposed schema for enhanced analytical capabilities. We will begin by focusing on key features such as NOC, Sport, Event, Year, Medal, Age, Height, and Weight. To answer our questions, we will employ various SQL functions, including COUNT(), AVG(), SUM(), and GROUP BY Clauses to aggregate data by NOC, Sport, Year, and Sex. For analyzing changes over time, we will use WHERE clauses to filter by specific year ranges and ORDER BY to observe trends. The primary metric for evaluation will be medal counts and averages of demographic data, allowing us to prove or disprove our hypotheses by identifying statistical relationships and trends within the Olympic data.