|  |  |  |
| --- | --- | --- |
| **StudentIDs:** | 266901 | Score: |
| **Class:** | Mini Project |  |
| **Date:** | 20-05-2024 |

This mini project assignment consists of 1 task. You should focus on providing complete solution, however if you are unable to solve a particular step, try to give at least a partial solution or provide justification for the reason for the lack of a solution.

# **Mini-project 1 – TOPic proposal**

The process of creating a data warehouse should be preceded by an understanding of the "business needs" and reality (problem field) represented by the available data resources. The implementation of the following task is to make you aware of the problems occurring in a specific (selected) section of reality, and then enable the identification (determination) of the needs, purpose, and capabilities of business analysis to support decision-making processes (making the right business decisions).

## **Task – details**

Please prepare the scope of the project in accordance with the specifications below, please remember that you should focus on one of the selected project datasets.

## **Task - Solutions**:

### 1.1 Project title

Analyzing Impact of Weather Conditions on Formula 1 Driver’s Race Performance

### 1.2 General description of the domain

Please introduce the selected domain.

Formula One (F1) racing, is one of the most prestigious and widely followed motor disciplines across the world. It is the premier class of international auto racing, governed by the Fédération Internationale de l'Automobile (FIA). Formula 1 consists of a set of races over different parts of the world on various international track layouts. All this is part of the annual championship, where both racers and constructors (i.e. race team) compete for points to determine the World Championship rankings. F1 is so much more than just a sport; it is the right blend of technology at its best, strategic planning, and real-time decision-making. Events generally span from race day, with practice and qualifying sessions to complete the schedule. The teams and drivers have to juggle a lot of variables, including how the car is performing, the driver's skill, track condition, and yes, weather. For this analysis we will limit the scope to only include data points from main F1 event and no qualification rounds etc.

### 1.3 Description of the analysis area with justification

Please state and comment the selected fragment of the domain intended for the data warehouse. Focus on identification and introduction of underlying business processes (events).

**Selected Fragment of the Domain**

The selected fragment of the domain for this data warehouse project is the analysis of the impact of weather conditions on Formula 1 (F1) drivers’ race performance during the main F1 races. This analysis focuses on integrating weather data with race performance metrics to provide actionable insights for F1 teams.

**Justification:**

* Strategic Planning and Race Strategy: Teams can use historical data on weather conditions and driver performance to devise strategies for future races. Understanding how weather impacts performance helps teams prepare better, choose appropriate tires, and optimize pit stop strategies.
* Performance Analysis and Improvement: Identifying patterns and correlations between weather conditions and driver performance enables teams to fine-tune their training and car setups. This helps in improving overall race performance under various weather conditions.
* Resource Allocation: Better understanding of weather impacts can help teams allocate resources more effectively, such as deploying specific engineering adjustments or focusing on driver training for adverse weather conditions.

**Business Processes (Events):**

* Race Events: The main race event, which determines the final positions, points, and overall performance metrics.
* Weather Events: Daily weather observations specific to the circuit location on the day of the main race.
* Performance Metrics: Metrics such as lap times, final positions, and points scored, analyzed in relation to the weather conditions on race day.

### 1.4 identified Problems

Please state and comment the identified decision problem(s) in the assumed domain.

**Decision Problem 1: Optimal Pit Stop Timing**

**Problem Statement:** Determining the optimal timing and frequency of pit stops based on historical weather data.

**Comment:** Historical data reveals patterns of effective pit stop timings under various weather conditions, helping teams plan pit stops to minimize time loss and maximize performance.

**Decision Problem 2: Driver Performance Trends**

**Problem Statement:** Understanding how different weather conditions have historically impacted driver performance and race outcomes.

**Comment:** Analyzing past races identifies trends in metrics like lap times and finishing positions under different weather scenarios, aiding teams in predicting performance and adjusting strategies.

**Decision Problem 3: Race Strategy Formulation**

**Problem Statement:** Formulating effective race strategies based on historical weather and performance data. **Comment:** Historical data on race strategies and outcomes under specific weather conditions helps teams develop robust future strategies, including driving styles, tire management, and pit stop planning.

**Decision Problem 4: Impact of Weather on Race Outcomes**

**Problem Statement:** Quantifying the impact of different weather conditions on race outcomes.

**Comment:** Statistical analysis of historical data shows how various weather conditions affect race results, allowing teams to prepare for adverse weather and optimize strategies for favorable conditions.

### 1.5 Project Goal

Please identify basic users (types) – provide brief description of each type – and state expectations of each type. Do not select too many types – as it might be beneficial to focus on more details for 3-4 user types. Further present detailed needs by specify 10-15 OLAP user query types (do not write any SQL queries, these should be formulated in natural language – strictly in business terms). Note – these should not be very specific queries (using specific attribute values), but rather general query/need types. For each query/need please indicate how a user can utilize resultant information. Match each query to the identified user types. Provide results in section 1.5.1.

Perform basic analysis of these query types (1.5.1), try to infer some general user requirements. Focus on identifying event(s) and perspectives that the user is interested in. Further, perform basic analysis of user query types (1.5.1), try to infer some detailed user requirements. List all possible measures and all possible individual dimensions. Provide results in section 1.5.2

#### 1.5.1 Expectations and detailed needs for decision support

User Types:

1. Race Strategists:
2. Description: Professionals responsible for planning race strategies, including pit stops and tire choices.
3. Expectations:

* Need to determine optimal tire selection based on weather forecasts.
* Require insights on the best pit stop times considering weather changes.
* Interested in historical performance trends under similar weather conditions.

1. Drivers:
2. Description: Athletes participating in races who need to adjust their driving based on weather conditions.
3. Expectation:

* Need information on how different weather conditions affect track grip and visibility.
* Require guidance on optimal driving styles and techniques for various weather scenarios.
* Interested in historical performance data of other drivers in similar conditions.

1. F1 Enthusiasts/Analysts:
2. Description: Fans and analysts who are interested in understanding the dynamics of the sport and predicting race outcomes.
3. Expectation:

* Want to understand the impact of weather on race outcomes for better predictions.
* Require visualizations of weather data and race results for analysis.
* Interested in correlations between weather conditions and specific race incidents.

OLAP User Query Types:

User 1: Race Strategists

* Query 1: "How does average lap duration vary across different weather conditions?"

Utilization: To adjust race strategies for upcoming races based on forecasted weather.

* Query 2: "What is the optimal pit stop duration under varying weather conditions?"

Utilization: To plan pit stops that minimize time loss.

* Query 3: "How do different constructors perform under the same weather conditions?"

Utilization: To assess competitive performance and strategize accordingly.

* Query 4: "What is the impact of weather changes on race performance for different circuits?"

Utilization: Helps in tailoring strategies specific to each circuit and expected weather.

* Query 5: "How do weather conditions affect the frequency of pit stops?"

Utilization: Guides on how often pit stops should be made during races with specific weather forecasts.

User 2: Driver

* Query 1: "What is the impact of visibility changes (e.g., fog, heavy rain) on lap times?"

Utilization: Helps drivers prepare for races with poor visibility conditions.

* Query 2: "What are the historical performance data of other drivers in similar weather conditions?"

Utilization: Allows drivers to benchmark their performance and learn from others.

Query 3: "What are the correlations between weather conditions and specific race metrics (e.g., fastest laps, pit stop durations)?"

Utilization: Enhances the understanding of how different weather conditions affect specific aspects of the race.

User 3: F1 Enthusiasts/Analysts

* Query 1: "How do weather conditions correlate with race outcomes (e.g., wins, podium finishes)?"

Utilization: Enhances understanding of weather impacts for better race predictions.

* Query 2: "What are the visualizations of weather data and corresponding race results for each season?"

Utilization: Provides comprehensive visual insights for analysis and reporting.

* Query 3: "How do weather conditions impact driver and constructor standings throughout the season?"

Utilization: Provides insights into how weather influences the overall championship standings.

* Query 4: "How do different weather conditions influence the performance of various constructors?"

Utilization: Helps in understanding which constructors perform better under certain weather conditions.

#### 1.5.2 Scope of analysis – aspects examined

1. General user requirements:

* Users are interested in the impact of weather on race events and performance.
* Key events include pit stops, race performance metrics, and overall race outcomes.
* Users need insights to improve strategic planning, performance evaluation, and safety measures.

1. Detailed user requirements:
2. Measures

* Average\_lap\_duration\_milliseconds: Average lap duration in milliseconds
* constructorsWins: Number of constructor wins
* driverWins: Number of driver wins
* fastestLap: Fastest lap time
* fastestLapRank: Rank of the fastest lap
* fastestLapSpeed: Speed of the fastest lap
* fastestLapTime: Time of the fastest lap
* laps: Number of laps
* Max\_pit\_stop\_duration: Maximum pit stop duration
* Median\_pit\_stop\_duration: Median pit stop duration
* milliseconds: total Duration in milliseconds
* Min\_pit\_stop\_duration: Minimum pit stop duration
* Min\_pit\_stop\_duration\_milisecons: Minimum pit stop duration in milliseconds

1. Dimension

A screenshot of a computer

Description automatically generated

### 1.6 Data sources

Prepare a brief description of the selected data source – focus on general structure and its usage, access options, update characteristics, volume, time span for facts, etc. Further perform initial assessment of the quality of available data – focus on number of records, number of valid records, number of missing records.

#### 1.6.1 Location, format, availability

**Selected Data Sources:** The data for this project is sourced from the Ergast API, which provides detailed historical race data, and the Visual Crossing Weather API, which provides historical weather data.

**Ergast API:**

* **Location:** Accessible online via the Ergast Developer API (http://ergast.com/mrd/)
* **Format:** Data is available in CSV format
* **Availability:** Freely available for public use with no access restrictions. Accessible vie RESTful API. Data can be queried using standard HTTP requests. Historical data is maintained and accessible.
* **Usage:** Ideal for integrating with analytical tools and platforms to extract and analyze F1 race data.

**Visual Crossing API:**

* **Location:** Accessible online via the Visual Crossing Weather API (https://www.visualcrossing.com/weather-api)
* **Format:** Data is available in JSON and CSV formats.
* **Availability:** Accessible vie RESTful API. Requires an API key, with free and paid tiers available.
* **Usage:** Suitable for obtaining detailed weather data to correlate with F1 race events for analytical purposes.

#### 1.6.2 Data source basic information

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Sr. num** | **Source** | **Table** | **Rows** | **Attributes** | **Size** | **Update Rate** | **Grain** |
| 1 | Ergast API | Circuits | 77 | 7 | 10 KB | Annually | Each record represents a single circuit. |
| Ergast API | Constructors | 212 | 4 | 18 KB | As new constructors enter the sport | Each record represents a single constructor. |
| Ergast API | Constructor Standings | 13212 | 5 | 306 KB | After each race | Each record represents a constructor's standings after a specific race. |
| Ergast API | Drivers | 859 | 7 | 92 KB | As new drivers enter the sport | Each record represents a single driver. |
| Ergast API | Driver Standings | 34470 | 5 | 853 KB | After each race | Each record represents a driver's standings after a specific race. |
| Ergast API | Lap Times | 568575 | 5 | 16,582 KB | After each race | Each record represents a single lap completed by a driver in a race. |
| Ergast API | Pit Stops | 10766 | 6 | 410 KB | After each race | Each record represents a single pit stop during a race. |
| Ergast API | Races | 1125 | 6 | 161 KB | Annually | Each record represents a single race event. |
| Ergast API | Results | 26400 | 14 | 1,654 KB | After each race | Each record represents a single driver's result in a race. |
| Ergast API | Status | 139 | 2 | 3 KB | As new status values are introduced | Each record represents a unique race status. |
| 2 | Visual Crossing API | Weather | 941 | 30 | 296KB | Daily updates | Each record represents weather for a specific date in specific location |

#### 1.6.3 Data source Initial assesment (Additional information. time span for facts, etc.)

**Data Quality Assessment:**

**Ergast API:**

* **Timespan:** Covers historical data from the inception of the Formula One World Championship in 1950 to the present.
* **Number of Valid Records:** Most records are valid, with comprehensive details for each race event.
* **Number of Missing Records:** Minimal missing records; some older historical data might be less detailed.

**Visual Crossing API:**

* **Timespan:** We are using the free tier for which historical weather data is available only from 1950 onwards.
* **Number of Valid Records:** High accuracy and completeness for recent data; some gaps may exist in older data.
* **Number of Missing Records:** Generally low, but varies by region and time period.

### 1.7-DIMENSIONAL model synopsis

Prepare a brief description of the hypothesized dimensional model – focus on underlying business process that your users are interested in (remember we are building a data mart – a single line of business is enough (one process) and required dimensions. Highlight the general structure, i.e., name facts and measures, name dimensions, and list required individual dimensions (dimension’s attributes).

**General Structure of schema:** The dimensional model is structured around the central fact table FactRacePerformance, which connects to various dimension tables via foreign keys. This schema structure resembles the snowflake schema and allows for efficient querying and analysis of the data.

**Here is how our hypothesized dimensional model looks like:**

A screenshot of a computer

Description automatically generated

#### 1.7.1 facts and measures

|  |  |  |  |
| --- | --- | --- | --- |
|  | Fact | Measure(s) | Grain |
| 1. | FactRacePerformance | Total\_time\_milliseconds  fastestLabRank  fastestLabTime  fastestLabSpeed  driverwins  constructorWins  Min\_pit\_stop\_duration  Max\_pit\_stop\_duration  Average\_pit\_stop\_duration  Average\_lap\_duration | Each record represents performance metrics for a driver in a specific race performance |

#### 1.7.2 Context for facts

|  |  |  |  |
| --- | --- | --- | --- |
|  | Dimension | Description | Grain |
| 1. | DimCircuit | Contains details about each circuit where races are held, including geographic location and names. | Each record represents a single circuit. |
| 2. | DimConstructor | Contains details about the constructors (teams) involved in F1 races, including their names and nationalities. | Each record represents a single constructor. |
| 3. | DimDate | Contains details about dates, including the day, month, and year, as well as more granular time details. | Each record represents a single date. |
| 4. | DimDriver | Contains details about the drivers participating in F1 races, including names, nationalities, and birthdates. | Each record represents a single driver. |
| 5. | DimPitstop | Contains details about pit stops during races, including durations and the number of stops. | Each record represents a single pit stop. |
| 6. | DimRace | Contains details about individual races, including the circuit, date, and round of the race. | Each record represents a single race event. |
| 7. | DimStatus | Contains various status values that can be assigned to race results, such as finished, accident, etc. | Each record represents a unique status. |
| 8. | DimWeather | Contains details about the weather conditions during the races, including cloud cover, temperature, and precipitation. | Each record represents weather conditions for a specific date and location. |

#### 1.7.3 Dimension overview

|  |  |  |
| --- | --- | --- |
|  | Dimension | Hypothesized Attributes |
| 1. | DimCircuit | circuitId, locationId, circuitRef, circuit\_name |
| 2. | DimConstructor | constructorId, constructorRef, name, nationality |
| 3. | DimDate | Date, Day, Month, MonthName, Year, time |
| 4. | DimDriver | driverId, driverRef, number, code, forename, surename, dob, nationality |
| 5. | DimLocation | locationId, country, location, lng, lat, alt |
| 6. | DimRace | raceId, circuitId, circuit\_name, round |
| 7. | DimStatus | statusId, status |
| 8. | DimWeather | weatherId, cloudcover, conditions, description, dew, feelslike, feelslikemax, feelslikemin, humidity, icon, precip, precipcover, preciptype, sealevelpressure, severerisk, snow, snowdepth, temp, tempmax, tempmin, visibility, winddir, windgust, windspeed |

### 1.7 Data source details

Study the selected dataset and present your findings in a form of a domain data dictionary (with information about table/sheet/location, attribute name, attribute type (high level type representation, like Numerical, Money, Text, Date, etc.), description (short description of the meaning of the attribute)) and quality assessment sheet (with information about table/sheet/location, attribute name, attribute type (lower-level type representation, like varchar(20), decimal(5,2), etc.), type of data (nominal, ordinal, interval, ratio, continuous), number of unique values, null ratio, quality assessment description (short description of the results of the attribute quality assessment – focusing on a column consistency assessment)).

### 1.7.1 Domain Data Dictionary

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | Location | Attribute name | Attribute type | Description |
| 1 |  |  |  |  |
| 2 |  |  |  |  |

### 1.7.2 Quality assesment Sheet

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Location | Attribute name | Attribute type | Type of data | Number of Unique values | Number of Null values | Quality assessment |
| 1 |  |  |  |  |  |  |  |
| 2 |  |  |  |  |  |  |  |

## **General Conclusions**:

Use this section to provide your general conclusions:

…

Remarks:

* A report without final conclusions will not be checked and results in a negative score!
* Quality, completeness, depth, and precision of provided descriptions affects your final score.