DS 2002 - Data Project 1
Midterm Project Documentation S
am Rea, Bryan Tompkins, Ben Alter

This document outlines the process of designing and executing an ETL (Extract, Transform, Load) pipeline for collecting data related to NFL games, teams, spreads, and weather information. The goal is to create a dataset that combines NFL game details with historical weather data so that we can query the data and learn about how weather affects NFL games and betting lines.

Data Sources:

1. NFL Game Data: We extracted NFL game data from Kaggle (https://www.kaggle.com/datasets/tobycrabtree/nfl-scores-and-betting-data/) containing tables with information about games, teams, spreads, and stadiums. These tables included data on game dates, game results, teams participating in each game, the associated point spreads and totals, the latitude and longitude of each stadium, the type of stadium (indoor/outdoor/retractable), and the location of the stadium.

2. Weather Data: We used a third-party weather data API, the Open-Meteo API, to retrieve historical weather data for each NFL game. This API provided information about max, mean, and min daily temperature and max daily wind speeds.

ETL Pipeline:

We implemented a Python-based ETL pipeline that performs the following operations:

1. Extraction: We extracted NFL game data from Kaggle. The data was stored in pandas DataFrames for further processing.
2. API Call: We made API requests to the Open-Meteo API to retrieve historical weather data for each game. The data was collected in JSON format.
3. Transformation: We combined data from various sources, mapping game data to corresponding weather information. Error handling was used.

Creating the Data Mart:

We designed a dimensional data mart that consists of a fact table table and three dimension tables:
1. NFL Games and Spreads - Fact Table
2. NFL Team information - Dimension Table
3. NFL Stadium information - Dimension Table
4. Weather Data - Dimension Table

These four tables are related by a common dimension, name of the nfl team which acts as a primary key in each of the tables.

SQL Queries:

We created SQL queries to illustrate the uses of the data mart.

1. SELECT Statements: We used SELECT statements to extract data from the four tables.

    a. SELECT * FROM weather ORDER BY numid LIMIT 5; b. SELECT * FROM teams ORDER BY numid LIMIT 5; c. SELECT * FROM stadiums ORDER BY sadium_id LIMIT 5; 2.

2. Aggregation: Aggregation functions such as the year the games were played, specific weather types and conference. were applied to demonstrate the capabilities of the data mart.

```
SELECT  DATE(game_date) AS game_date, COUNT(*) AS games_played
FROM  nfl_game WHERE temperature < 32
GROUP BY DATE(game_date)
```

Conclusion:

In conclusion, we successfully designed and executed an ETL pipeline to collect NFL game data, teams, spreads, and historical weather information. This dataset can now be used for analysis to understand the impact of weather conditions on NFL games and spreads.