CSI4142 Fundamentals of Data Science Project Phase 1: Conceptual Design – Dimensional Model

Group 1

Professor Yazan Otoum,

TA: Lansu Dai

Jenson Wu #300166874
Victor Feng #300176400
Lootii Kiri #300189957

Date: February 1st 2024
Due Date: February 9th 2024

Grain:
Our focus is on individual flights from a specific airline departing from a particular airport on a set date, bound for a specific destination. We aim to determine if these flights are canceled or if delayed, to identify the type of delay and quantify the total number of delays.

Trends:
The purpose of this project is to identify the root causes of flight distribution. We have gathered 5 of the most common causes (weather, carrier, security, traffic and late aircraft) to see if there is a recurring trend in certain parts of the country. We have also gathered a list of aircrafts, to analyze if the model and/or manufacturer has any impact on their flight performance.

Key Indicator:
Delayed and canceled flights are what causes the most distribution. In our fact table, we want to track whether the flight was canceled as well as if there were any delays to the flight according to the constraints that we used.

Fact Table

Attributes	Type	Min	Max	Sample
Is Canceled	Boolean	False	True	True
Weather Delay	Integer	0	10000	310
Carrier Delay	Integer	0	10000	1781
Security Delay	Integer	0	10000	16
Traffic Delay	Integer	0	10000	281
Late Aircraft Delay	Integer	0	10000	0

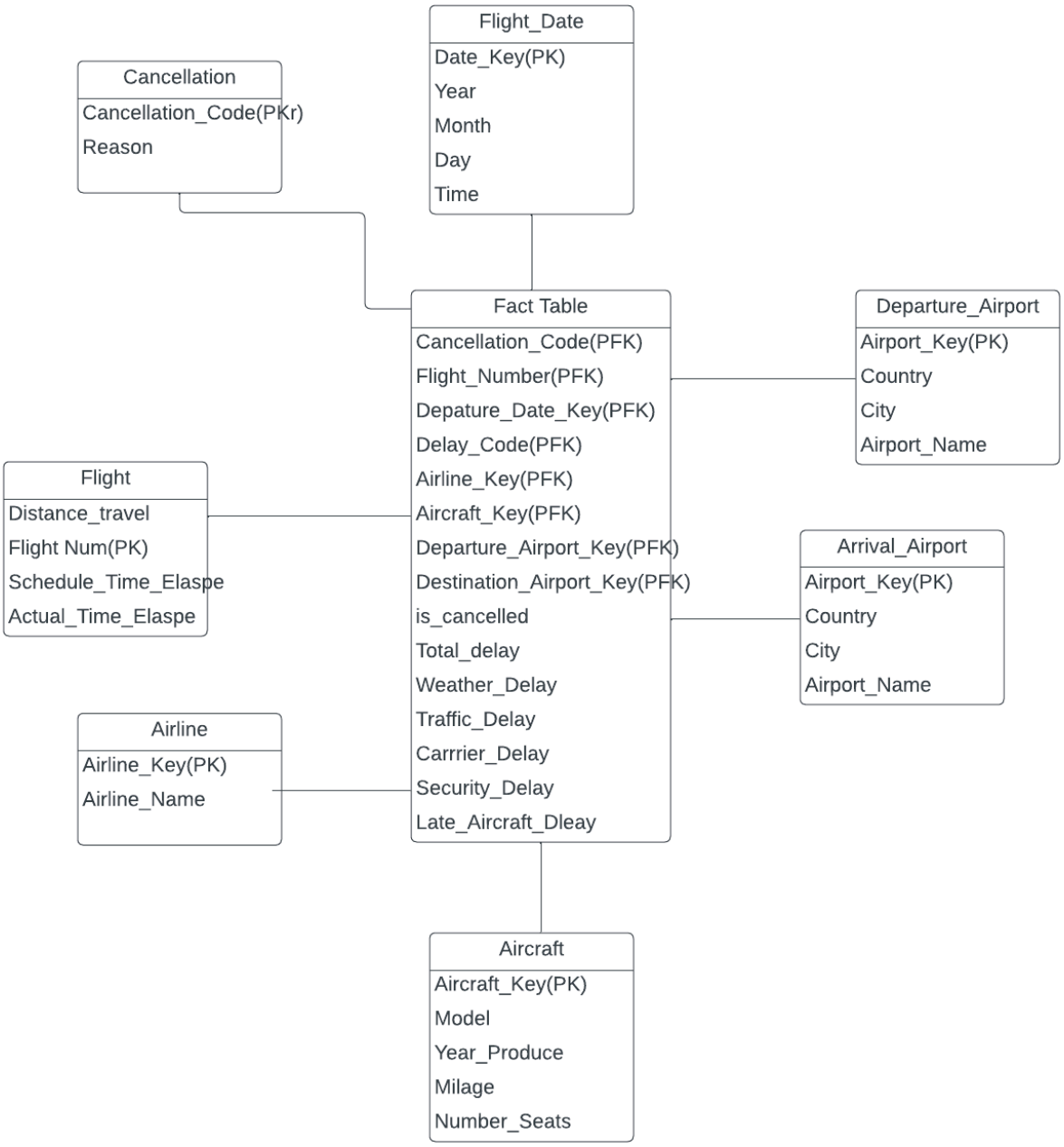
Dimensions:

Flight				
Attributes	Type	Min	Max	Sample
Flight Number(Primary	Integer	0	100000000000	1901

Flight				
Attributes	Type	Min	Max	Sample
Key)				
Estimated flight time (minutes)	Integer	0	100000	2901
Actual flight time elapsed (minutes)	Integer	0	100000	181
Flight Distance (Kilometer)	Integer	0	100000	213
Airline				
Airline ID (PK)	Integer	0	100000	123
Airline Name	String			“Delta”
Aircraft				
Aircraft ID (Primary Key)	Integer	0	100000	1010
Model	String			“Boeing 747”
Milage	Integer	0	100000	1010
Number of seats	Integer	0	100000	12191
Arrival Airport				
Airport ID(Primary Key)	Integer	0	1000000	11811
Airport Name	String			“Denver International Airport”
Country	String			“United States”
City	String			“Denver”
Destination Airport				
Airport ID(Primary Key)	Integer	0	1000000	811
Airport Name	String			“Ottawa International

Flight				
Attributes	Type	Min	Max	Sample
				Airport”
Country	String			“Canada”
City	String			“Ottawa”
Departure Date				
Date_Key(Primary Key)	Integer	0	1000	29
Year	Integer	2010	2024	2013
Month	Integer	1	12	2
Day	Integer	1	30	18
Flight Date				
Date_Key(Primary Key)	Integer	0	100000	2012
Year	Integer	2010	2024	2018
Month	Integer	1	12	6
Day	Integer	1	31	24
Cancellation				
Cancellation Code(Primary Key)	Integer	0	4	1
Reason	String			“Weather”

Conceptual Model



Assumptions:

- Every flight that is canceled must have a reason for cancellation
- Multiple types of delays can be applied to the same flight
- No flight are diverted to another airport that's not the original arrival airport
- All flight data are correct and no need to overwrite

Design Mistakes

	Design Mistake	Implementation to avoid design mistakes
1	Avoid placing text attributes in the Fact table.	In our fact table, instead of having the type of delay(string attribute), we have an individual attribute for each type of delay that may occur during a flight.
2	Avoid limit verbose descriptions to save space.	In our fact table, we have look up tables that stores the more verbose descriptive attributes
3	Avoid Normalise to save space (leads to slower queries!).	In our conceptual model, we use string attributes for dimension table to create a star schema and avoid a snowflake schema
4	Avoid Ignoring the need to track changes.	In our conceptual model, to ignore slowly changing dimensions, we will use SCD Type 2-C (includes the date for each change made).
5	Add new hardware to solve all query performance issues	We will use a cloud server to store all data within our data mart. This optimizes querying for data retrieval efficiently within our existing infrastructure and without needing to add new hardware.
6	Use operational key's as the primary keys.	We used the surrogate key as the primary key in our conceptual model. For instance, the tail number can be used as the primary key, but we still use the surrogate key for that.
7	Avoid neglecting to declare (and comply with) the grain.	Our grain reflects information we have in our fact table, we got our grain from analyzing business processes, and use it to design our fact table and attributes.
8	Avoid neglecting a detailed design.	We used two different(Flight and aircraft dataset) to enrich our data mart
9	Avoid expecting users to query normalized data	We will develop a BI dashboard that allows user to filter and interact with the data without creating their own queries
10	Avoid failing to conform to Facts and Dimensions.	To avoid failing to conform facts and dimensions, we ensured that all facts in our data mart are consistently linked to appropriate dimensions, maintaining data integrity and facilitating accurate analysis.

Work Distribution:

Student Number	Name	Task
300166874	Jenson Wu	Discuss and Design Dimension and Fact Table , Design mistakes
300176400	Victor Feng	Discuss and Design fact table and Dimensions, Design mistakes.
300189957	Lootii Kiri	Grain, Design mistakes
	Group	Conceptual Model

Note: We meet with the TA on a weekly basis, where we discuss the progress we have made thus far.

Sources:
Flight Dataset: Download page (bts.gov)
Aircraft Dataset: Aircraft Dataset