Washington State University
CPT_S 415 – Big Data
Online

Srinivasulu Badri

Assignment 1

Name: Nam Jun Lee

Student Number: 11606459

1. [**Big Data concept**] Give one example of Big Data application you know. Use the detailed example to explain each of the five Big V’s.

The big data application I know is Facebook. So let me explain how 5V’s of big dat a are applied on Facebook.

Five Big V's:

(1) Volume: Facebook has a huge number of users. According to Google, monthly active users are 2.934 billion. As such, the amount of data Facebook has is enormous.
(2) Velocity: Facebook speeds up scrolling posts and playing videos, so users don't feel uncomfortable using the service
(3) Variety: Facebook contains text such as users' personal information and messag es. There are also pictures and countless videos. As such, Facebook includes v arious data source types.
(4) Veracity: Facebook contains the quality and accuracy of its data. They remove dirty data for the smooth operation of the service so that users do not feel un pleasant emotions while using the service.
(5) Value: Facebook includes a 'Like' feature. Through the information obtained by this function, the product advertisement for each user's area of interest is disp layed intensively, thereby arousing the desire of users to purchase. As such, it helps to create business value by analyzing users' interests.

2. [**Relational Data Model**] As of January 2017, the OpenFlights Airports Databa se (https://openflights.org/data.html) contains over 10,000 airports, train stations and ferry terminals spanning the globe. Each entry in the Airport table contain s the following:

    A. Consider the following terms: *relation schema, relational database sche ma, domain, attribute, attribute domain, relation instance.* Give what the se terms are with the above Airport database. Give one small (4-5 tuple s) instance of the Airport table.

Airport

Airport ID	Name	City	Country	IATA	ICAO	Latitude	Longitude	Altitude	Time zone	DST	Tz database time zone	Type	Source
1	Goroka Airport	Goroka	Papua New Guinea	GKA	AYGA	-6.08168983459 0001	145.39199829 1	5282	10	U	Pacific/Port_Moresby	airport	OurAirports
2	Madang Airport	Madang	Papua New Guinea	MAG	AYMD	-5.20707988739	145.78900146 5	20	10	U	Pacific/Port_Moresby	airport	OurAirports
3	Mount Hagen Kagamuga Airport	Mount Hagen	Papua New Guinea	HGU	AYMH	-5.826789855957031	144.29600524902344	5388	10	U	Pacific/Port_Moresby	airport	OurAirports
4	Nadzab Airport	Nadzab	Papua New Guinea	LAE	AYNZ	-6.569803	146.725977	239	10	U	Pacific/Port_Moresby	airport	OurAirports

Relation schema: Airport(AirportID, Name, City, Country, IATA, ICAO, Latitude, Longitude, Altitude, Time zone, Tz database time zone, Type, Source)

Relational database schema: Database schema is the collection of relation schema. Hence, one relational database schema.

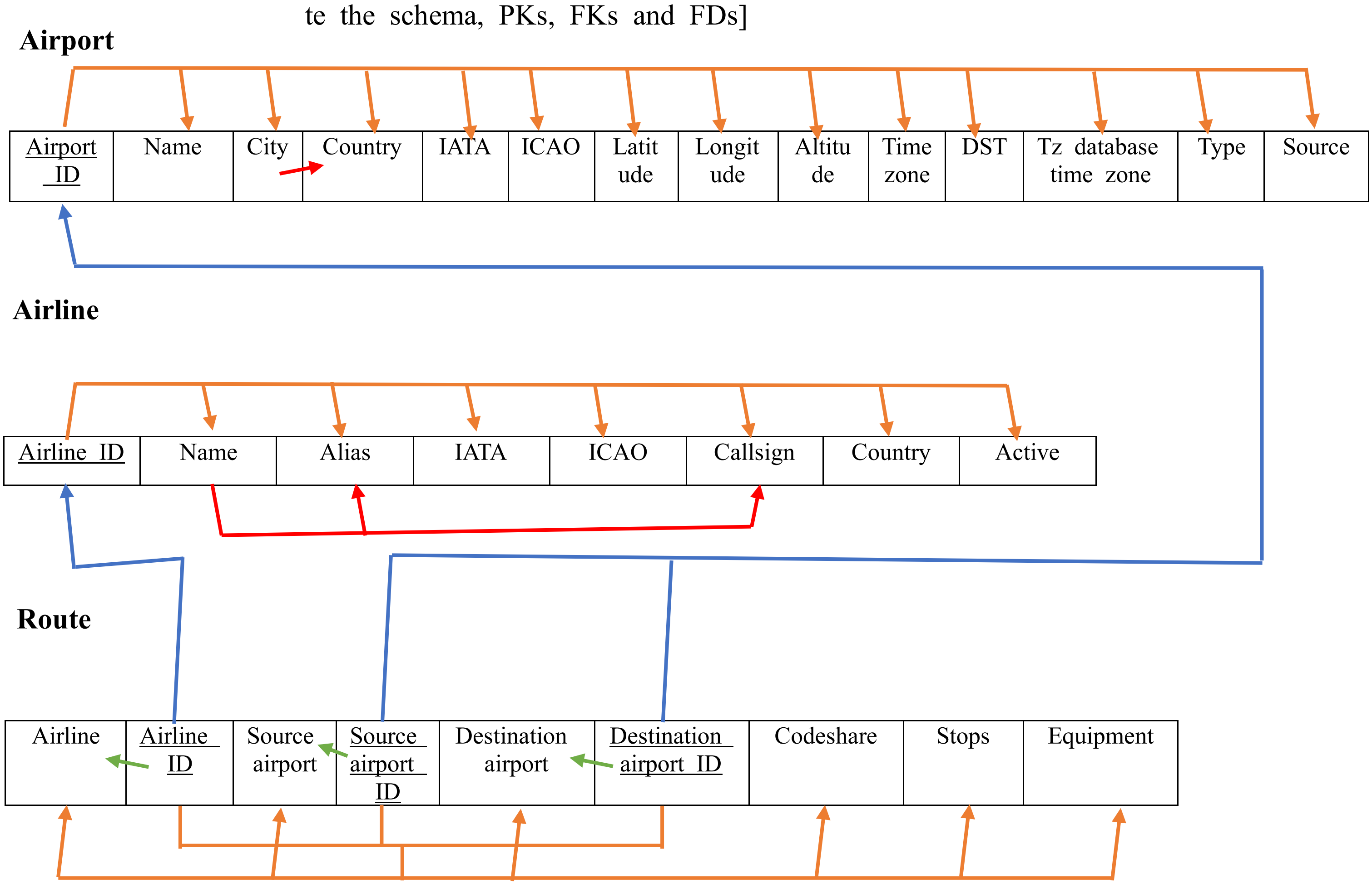
Domain: A data type with optional constraints. For example, the domain of Name, City, and Country in the Airport relation should be a string. It cannot have numeric values.

Attribute: Column Headers. Hence, Airport table has 14 attributes: AirportID, Name, City, Country, IATA, ICAO, Latitude, Longitude, Altitude, Time zone, Tz database time zone, Type, Source.

Attribute domain: All possible values in a Column. For example, the domain of IATA in the Airport relation is 3-letter IATA code.

Relation instance: A relation instance is a tuple/row in a relation. Hence, the Airport table has one particular combination of attribute values.

B. There are three databases in the OpenFlight dataset: Airport, Airline, an d Route. Give the schema of these three databases and mark the primar y keys, foreign keys and provide examples of functional dependencies y ou identified over the three tables. [You may draw a diagram to illustra te the schema, PKs, FKs and FDs]



Primary Key: Each schema is underlined (____)

Foreign Key: Directed arrow from FK to referenced schema ( )

Functional Dependency: Directed arrow from FD to related attributes ( )

Partial Dependency: Directed arrow from partial dependency to related attributes ( )

Transitive Dependency: Directed arrow from transitive dependency to related attributes ( )

3. [**Functional Dependencies**] Recall Armstrong's axioms.

- Reflexivity rule: if $Y \subseteq X$ then $X \rightarrow Y$
- Augmentation rule: if $X \rightarrow Y$ then $XZ \rightarrow YZ$
- Transitivity rule: if $X \rightarrow Y$ and $Y \rightarrow Z$ then $X \rightarrow Z$

a. Give two examples for using Armstrong's inference rules to induce new FDs from the set of FDs you designed in question 2 (b).

(1) Augmentation rule: Airport Schema

If Airport ID $\rightarrow$ Country then,

Airport ID + Name = Country + Name

(2) Transitivity rule: Airline Schema

If Alias $\rightarrow$ Name and Name $\rightarrow$ Callsign then,

Alias $\rightarrow$ Callsign

b. Prove the following inference rules also hold, using FD definition and Armstrong's Axioms.

i. decomposition rule: **if $X \rightarrow YZ$ then: $X \rightarrow Y$ and $X \rightarrow Z$**

Proof:

$X \rightarrow YZ$ ----- given above

$YZ \rightarrow Y$ ----- Reflexivity Rule

$YZ \rightarrow Z$ ----- Reflexivity Rule

$X \rightarrow Y$ ----- Transitivity Rule on $X \rightarrow YZ$ and $YZ \rightarrow Y$

$X \rightarrow Z$ ----- Transitivity Rule on $X \rightarrow YZ$ and $YZ \rightarrow Z$

Hence, Proved.

ii. Pseudo transitivity: **if $X \rightarrow Y$ and $YW \rightarrow Z$ then: $XW \rightarrow Z$**

Proof:

$X \rightarrow Y$ ----- given above

$YW \rightarrow Z$ ----- given above

$XW \rightarrow YW$ ----- Augmentation Rule on $X \rightarrow Y$ by augmenting with W

XW → Z ----- Transitivity Rule on XW → YW and YW → Z

Hence, Proved.

4. [**Normalization]** Given a relation R(A_1, A_2, A_3, A_4), with three FDs A_2, A_3 → A_4; A_3, A_4 → A_1; A_1, A_2→ A_3. Provide the 3NF and BCNF form of the schema and explain why.

First, find the candidate keys.

Check FDs: A_1, A_3, A_4 are included in both attributes, but A_2 is included only in the left attribute. So, calculate the closing of the power set combination of A_2:

$A_1, A_2 = A_1, A_2, A_3, A_4 = R$ ---- Candidate key

$A_2, A_3 = A_1, A_2, A_3, A_4 = R$ ---- Candidate key

$A_2, A_4 = A_2, A_4 \subset R$

$\{A_1, A_2\}, \{A_2, A_3\}$ lead only to superkeys, so $\{A_1, A_2\}, \{A_2, A_3\}$ are candidate keys.

Hence, the Set of key attributes: A_1, A_2, A_3.

Now, provide the 3NF and BCNF form of this relation R:

- 3NF: For each FD, determine whether the left-hand side is a super key or the right-hand side is all key attributes, **this relation is in 3NF** because there is no transitive dependency here.
- BCNF: Check each FD: $\{A_3, A_4\}$ is non-trivial but is not a super key. If this relation is in BCNF, each non-trivial FD must be a super key. Hence, **it violates BCNF**. It means **this relation is not in BCNF.**

.