Airline Data Warehouse

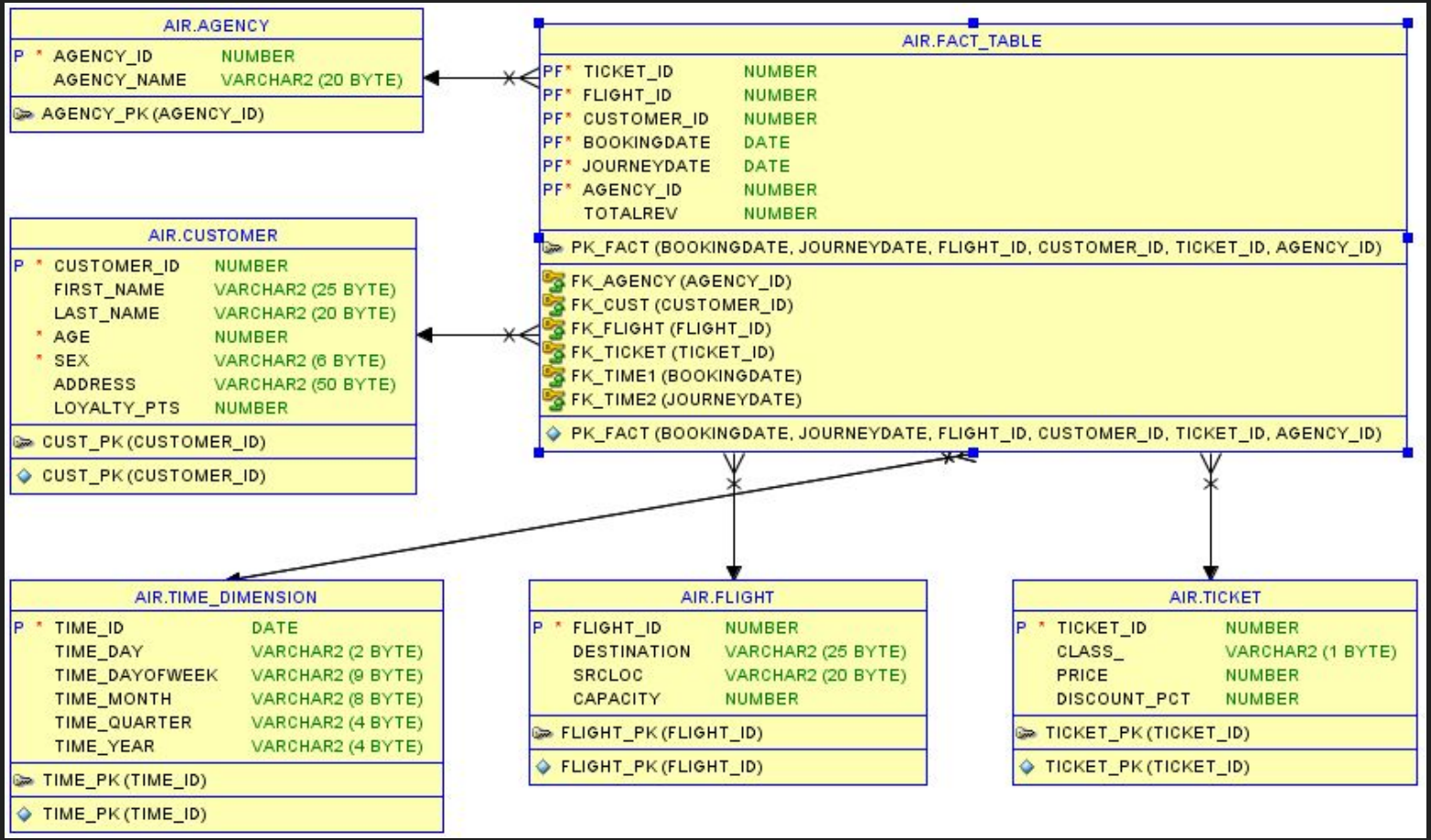
Dhanur Sharma
Nahush Bhobe
Aadhar Gautam

Problem Statement

To design a data warehouse for an airline to be able to answer key business questions and provide optimal solutions for pertinent issues.

We will be using 5 dimensions, namely, Time, Ticket, Flight, Agent, and Customer, and arranging it in a star model connected to the Fact table.

Relational Model



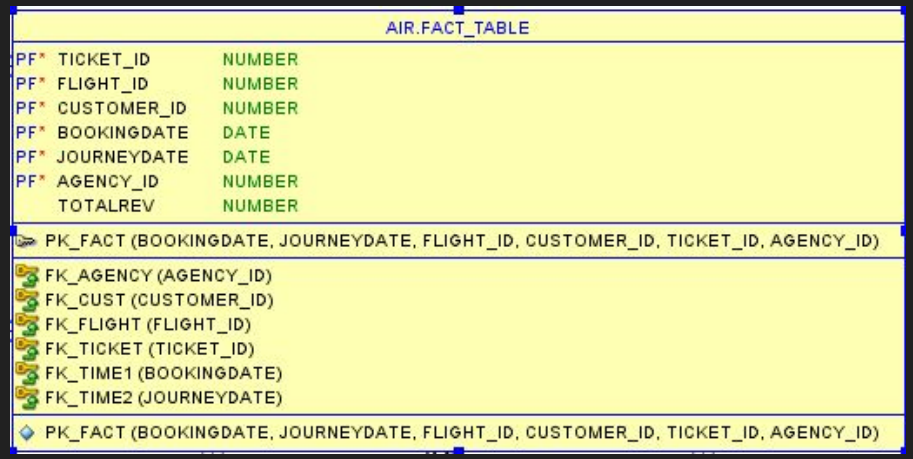
Fact table

Inherits the following Primary keys from the Dimension tables:

- flight_id
- customer_id
- ticket_id
- bookingdate
- journeydate
- agency_id

Measures used:

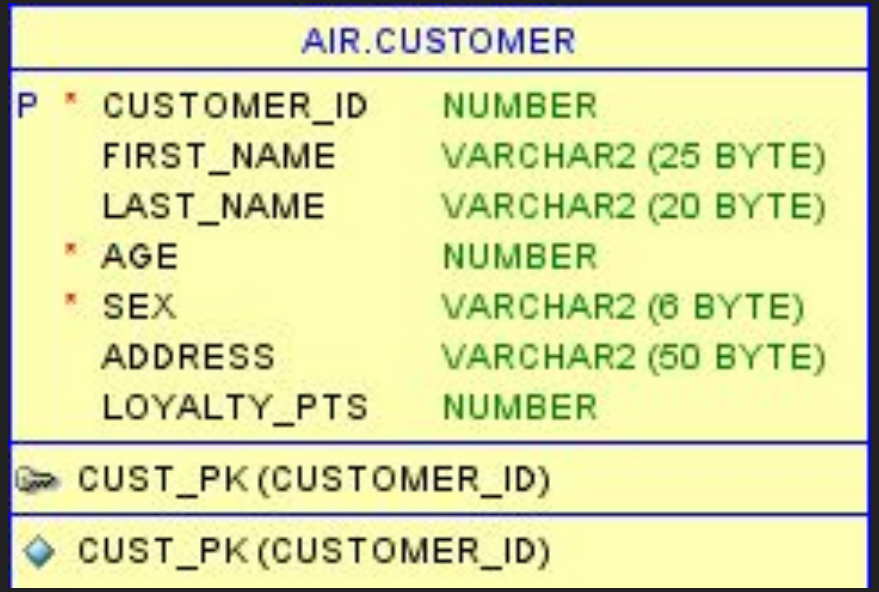
total_revenue



Customer table

Contains the following attributes:

- cust_id
- first_name
- last_name
- age
- sex
- address
- loyalty_pts



Flight table

Contains the following attributes:

- flight_id
- destination
- srcloc (source)
- capacity

AIR.FLIGHT		
P *	FLIGHT_ID	NUMBER
	DESTINATION	VARCHAR2 (25 BYTE)
	SRCLOC	VARCHAR2 (20 BYTE)
	CAPACITY	NUMBER
FLIGHT_PK (FLIGHT_ID)		
FLIGHT_PK (FLIGHT_ID)		

Ticket table

Contains the following attributes:

- ticket_id
- class_
- price
- discount_pct

AIR.TICKET		
P *	TICKET_ID	NUMBER
	CLASS_	VARCHAR2 (1 BYTE)
	PRICE	NUMBER
	DISCOUNT_PCT	NUMBER
TICKET_PK (TICKET_ID)		
TICKET_PK (TICKET_ID)		

Time table

Contains the following attributes:

- time_id
- time_day
- time_dayofweek
- time_month
- time_quarter
- time_year

AIR.TIME_DIMENSION		
P *	TIME_ID	DATE
	TIME_DAY	VARCHAR2 (2 BYTE)
	TIME_DAYOFWEEK	VARCHAR2 (9 BYTE)
	TIME_MONTH	VARCHAR2 (8 BYTE)
	TIME_QUARTER	VARCHAR2 (4 BYTE)
	TIME_YEAR	VARCHAR2 (4 BYTE)
TIME_PK (TIME_ID)		
TIME_PK (TIME_ID)		

Agent table

Contains the following attributes:

- agency_id
- agent_name

AIR.AGENCY		
P *	AGENCY_ID	NUMBER
	AGENCY_NAME	VARCHAR2 (20 BYTE)
AGENCY_PK (AGENCY_ID)		

Benefits of using a Data warehouse

- The airline industry runs multiple types of operations simultaneously, such as customer service, baggage handling, flight scheduling, ticket sales, and overall business management. As a result, airlines collect and store large amount of heterogeneous data from a wide variety of sources that they can leverage to identify opportunities to improve processes, reduce costs, and increase revenues.
- Data warehousing provides a centralized repository for corporate data and information assets. A data warehouse is not identical to the organization’s database used for transaction processing.
- This process of centralized data management and retrieval rely on data warehouses, which is defined as a subject-oriented, integrated, time-variant, nonvolatile collection of data in support of management's decision making process.

- The company can leverage data mining techniques to improve customers' loyalty through market segmentation, understand what their competitors are doing, forecast sales, monitor business performance, and detect fraud, waste, and abuse.
- It also allows organizations to analyze data from multiple perspectives, categorize it, and uses the information to predict future trends and behaviors, decrease costs, increase revenues, and improve processes.
- The architecture is open and scalable and built in such ways that it can support the future expansion of data.

Challenges faced while creating a data warehouse

- A key implementation challenge for data warehouses is integrating conflicting or redundant data from different sources. Furthermore, the size of the database and query complexity will affect the type of system needed by organizations.
- The implementation of a data warehouse is often a long-term, time-consuming, and resource intensive process. Organizations may not have the necessary expertise to setup and maintain a data warehouse or they may over-estimate the needs of the system, thus leading to higher costs.
- Data warehouses capture only a fraction of the information needed by managers for decision-making activities and they often cannot collect, retrieve, and disperse worker’s knowledge.

Challenges in the Airline industry for data warehousing

- The airlines industry faces interoperability issues as it uses multiple and complex information technology (IT) systems to support their operations.
- Many airlines operate older legacy systems, which can create IT systems integration issues and make it difficult to deploy data warehouses and data mining software.
- Multiple mergers and acquisition can also complicate the integration and reconciliation of conflicting or redundant data and can comprise data integrity and security.
- Implementing a data warehouse can be costly and cash- strapped airlines may be unable to secure the necessary funding.
- Another challenge for airlines as they collect information on passengers is the need to balance the privacy of their customers with the requirements of providing government agencies with the necessary information to support national security efforts.
- Airlines must be able to exchange data and information with their multiple business partners, thus complicating further the need for IT systems integration.

Success story in the airline Industry to further endorse the implementation of data warehousing

Continental Airlines (Continental) ranked the lowest among major U.S. Airlines in regards to on-time performance, mishandled baggage, and customer complaints. Historically, Continental had outsourced its operational systems and only received a limited set of scheduled reports and no support for ad hoc queries.Continental decided to develop an enterprise data warehouse, which the CIO identified as “core to Continental strategy and thus should not be outsourced”.The real-time information came from multiple sources including the mainframe reservation system, satellite feeds from airplanes, and a central customer database. The airline leveraged the real-time information to improve the recovery of lost airline reservations, customer value analysis, marketing insight, flight management dashboard, and fraud investigations. Over six years, Continental invested US $30 million in hardware and software that realized over US $500 million in increased revenues and cost savings, and went from “worst to first”.

Sample queries

```sql
--revenues per month
SELECT t.time_month,SUM(totalrev) from fact_table f, time_dimension t WHERE f.bookingdate = t.time_id GROUP BY t.time_month;
```

	TIME_MONTH	SUM(TOTALREV)
1	DEC	768253024.79
2	NOV	741225883.52
3	MAR	769197121.69
4	APR	745596324.78
5	SEP	738765568.86
6	OCT	766915590.99
7	FEB	694459606.37
8	AUG	769659371.86
9	JUL	763122883.38
10	JUN	745948349.4
11	MAY	768956966.65
12	JAN	773380289

Sample queries

```
--revenues per month and year
SELECT t.time_year,t.time_month,SUM(totalrev) FROM fact_table f, time_dimension t
WHERE f.bookingdate = t.time_id GROUP BY t.time_month,t.time_year ORDER BY t.time_year;
```

Query Result

All Rows Fetched: 25 in 1.373 seconds

	TIME_YEAR	TIME_MONTH	SUM(TOTALREV)
1	2018	APR	371240537.85
2	2018	AUG	385094294.34
3	2018	DEC	390243091.05
4	2018	FEB	347509898.09
5	2018	JAN	339707361.45
6	2018	JUL	379753221.73
7	2018	JUN	374122441.19
8	2018	MAR	382848714.67
9	2018	MAY	385048599.8
10	2018	NOV	371474832.24
11	2018	OCT	382660778.93
12	2018	SEP	367744402.39
13	2019	APR	374355786.93
14	2019	AUG	384565077.52
15	2019	DEC	378009933.74
16	2019	FEB	346949708.28
17	2019	JAN	388901787.76
18	2019	JUL	383369661.65
19	2019	JUN	371825908.21
20	2019	MAR	386348407.02

Sample queries

```
227 --number of tickets between 2 locations
228 SELECT COUNT(ticket_id) FROM fact_table WHERE flight_id = (SELECT flight_id
229                                                            FROM flight
230                                                            WHERE destination = '&in1'
231                                                            AND srcloc = '&in2');
```

Query Result

SQL | All Rows Fetched: 1 in 0.078 seconds

	COUNT(TICKET_ID)
1	25193

Sample queries

```
232 --number of tickets for a specific agency code
233 SELECT COUNT(ticket_id) FROM fact_table WHERE agency_id = '&agent_id';
```

Query Result

SQL | All Rows Fetched: 1 in 0.078 seconds

	COUNT(TICKET_ID)
1	133495

Sample queries

```
--number of passengers in their classes
SELECT t.class_,COUNT(customer_id) FROM fact_table f, ticket t WHERE f.ticket_id = t.ticket_id GROUP BY t.class_;
```

Query Result

SQL | All Rows Fetched: 2 in 0.826 seconds

	CLASS_	COUNT(CUSTOMER_ID)
1	B	266565
2	E	533435

Thank you!