**Data Warehouse Project**

**Compass Airlines**

**By Anbu Malar, Deepti Kulkarni, Hoang Ho,  Reshma Paturi, Shivani Deshpande**

**a.       Short overview of business**

Compass Airlines is a regional airline with hubs in Los Angeles international airport (LAX), Phoenix Sky Harbor international airport (PHX), and Seattle-Tacoma international airport (SEA). It currently is operating 56 aircrafts and mostly serving as regional carriers for Delta Airlines (under the name of Delta Connection) and American Airlines (under the name of American Eagle).

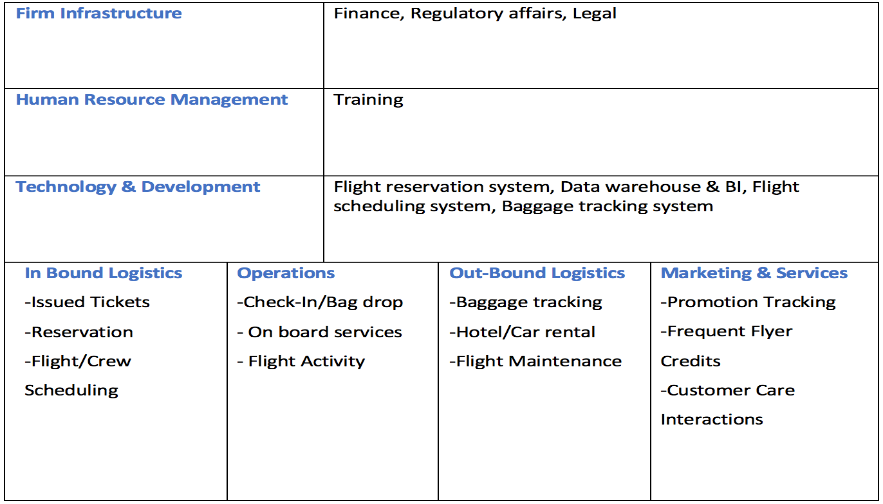
As any other regional airlines, the main business activities of Compass is transporting passengers between regional cities, as its fleet consists of 100% Embraer 175 aircraft whose service range is around 1,750 nautical miles. Now, Compass Airlines serves over 40 destinations, with a concentrated coverage on the West coast. Main business processes associated with the company are ticket issuance, flight/crew scheduling, aircraft maintenance, baggage tracking, and customer-related support. These processes will be discussed further in detail in the sections below.

**b.      Overview of Data that includes**

**(i) Transactional data in relational databases:**

Airline industry deals with several processes simultaneously, such as flight operation, customer care, flight reservations, baggage tracking, and aircraft maintenance. All these processes generate great amount of data, both structured and unstructured.

The value chain of the airline industry is as follows:



**Transaction data in Relational database:**

Aviation industry is long known for putting transactional data to good use through analytics.

Few such examples are:

* Controlling the airfare by the customer behavior, route traffic history.
* Seasonal price/demand prediction for a route.
* Air traffic management.
* Promotions to cater to a customer’s profile.

The transactional data are collected during various steps of the value chain. For instance:

**Human Resources Management:**

* Crew profile
* Other faculties profiles
* Training coursework, certifications
* Trainers, school details

**In-Bound Logistics:**

The inbound logistics involves route selection, flight reservation, flight maintenance, and flight/crew schedule.

**Flight Reservation system:**

This is the most significant revenue source of the company and is a very important area to perform analytics on.

The flight reservation system is where the customers book flight tickets. This includes third-party websites and physical ticketing office. The data collected here is crucial to study the customer behaviors.

Flight reservation systems include processes such as issued tickets and reservations. The data collected here is:

* + PNR number
  + Customer profile
  + Route chosen
  + Origin, destination
  + Promotions
  + Webpage
  + Change details (change data, cancellation, change destination)
  + Seat selection
  + Baggage
  + Insurance

**Flight/Crew Scheduling System:**

Every flight segment is pre-planned and coordinated with the airports much ahead for efficient performance.

The data collected here would include:

* + Crew members schedules
  + Flight schedules
  + Flight details
  + Airport details (origin and destination airports)
  + Traffic details issued by each airport’s air traffic management

**Operations:**

**Baggage Tracking:**

This involves baggage tracking and details to reduce mishandling and to ensure the baggage are delivered on time at the right destinations. Analytics can particularly be useful in identifying defaults and quality maintenance.

* + Check in baggage details (weight, smart tag)
  + Passenger travel details
  + Smart baggage tracking

**Flight Activity:**

Check-in/Bag drop services:

This includes data from the check-in counter and gates. The data collected here includes:

* + Passenger boarding details
  + Baggage weight and specifications
  + Any changes to the reservation
  + Payment for extra services

On-board services:

There are shopping and food services available in each flight.

* + Inventory details for each flight segment
  + Third-party (food caterer) details
  + Shopping order

**Out-Bound Logistics:**

**Rental Services:**

* Third Party profile (hotels, car rental services)
* Bookings
* Discounts for services

**Aircraft Maintenance:**

* Third-party profile (fuel service provider, airport services)
* Quality check logs
* Maintenance team profile

**Marketing:**

**Promotion tracking: Frequent flyer credit:**

* Advertisements
* Promotions details (route, period)
* Promotion order details
* Customer profile
* Mileage credit history
* Route details (flight, airport origin, destinations)
* Order detail

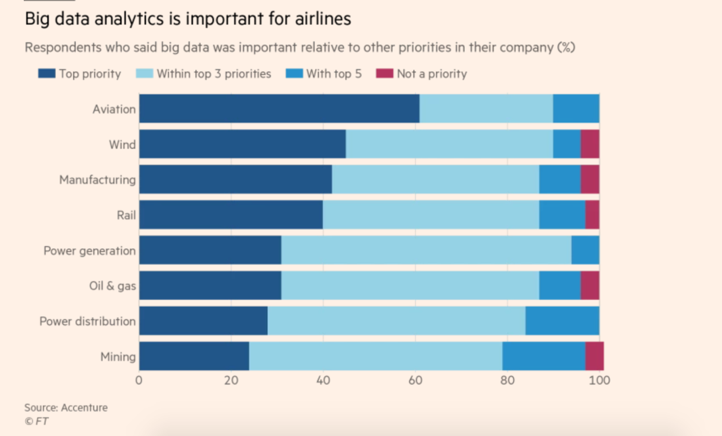
**Services:**

**Customer Care Interactions:**

* Customer care staff profiles
* Customer care physical store details
* Baggage delay complaints
* Insurance transactions
* Complaint/feedback
* Customer profile

**2. Big data from Social and/or Sensor or other sources.**

Airline industry deals with vast amount of big data in their daily operations. Since this industry is heavily competitive, it is important to stay ahead in the market. Data warehousing and data mining can play a key role in decision making. A study from Accenture states that big data analytics has become the highest priority for aviation (61%) compared other industries. The graph below depicts how important big data analytics is perceived to be. Airline industry heavily relies on customer satisfaction for better profit margin. Emerging big data analytics can be greatly useful in studying customer behaviors to develop more suitable marketing strategies, which include setting appropriate price range and personalizing offers tailored to customer needs.

****

(Source:<https://www.ft.com/content/f3a931be-47aa-11e8-8ae9-4b5ddcca99b3>)

The company collects unstructured big data from several business process, from the data collected through reservation systems to IoT data from sensors in the flight. A study by FAA states that during a year, on an average, a jet engine generates up to 10 TB of data. Most of these data are as unstructured big data.

Some of sources of big data are as follows:

|  |  |
| --- | --- |
| **Business Process** | **Big Data Source** |
| Ticket Reservation | Type of tickets searched for (refundable, one- way), clickstream data Number of times the ticket is searched from an IP before booking (‘Look to book’)  Third party ticket cost fluctuations over time. Time and place of flight reservation |
| Customer Care Interactions | Customer care audios, emails, feedback forms, customer information from check-in desks, support centers |
| Baggage tracking | History of real-time baggage data, baggage delay feedback, insurance transactions, RFID and IoT details of baggage |
| Promotion Tracking | Social media data such as tweets, likes for promotions, news, clickstream data; advertisements on third party websites, emails. Physical promotions, discounts, personalized offers |
| Flight Activity | Sensor data in the entire flight, space utilization on flights, aircraft movement, communications, air traffic management data, reported incidents |
| Aircraft Maintenance | Business transactions with the maintenance team, sensor data, quality check reports |
| Flight/Crew Scheduling | History of crew schedule for all the flights, crew incidents, feedback |
| Flight tracking | Current flight status data, flight status data of competitors (third-party data such as FlightAware) |
| Third-party data | Meteorological data, IoT data from the airports where flights originate/terminate; air traffic data, hotel/car rental feedback, and surveillance data |

**c.       Enterprise Data Warehouse Bus Architecture for the Mid-Size Business**

The figure below shows Enterprise Warehouse Bus Matrix for Compass Airlines. Following are the dimensions and business processes associated with the company:

**Conformed Dimensions:**

1. **Airport**: information of the name of different airports compass airlines has stops at and all other details like airport number, name, location, etc.

2. **Aircraf**t: details of each individual Compass aircrafts including aircraft no, type of aircraft ( 77, 797 etc.), capacity per flight, etc.

3.  **Passenger**: contains the details of each passenger such as name, phone no, address, zip code, etc.

4. **Date**: important dates from booking and scheduling perspective. Has Hierarchy year month date

5. **Time**: total time and represents exact position of the flight in journey and timestamp for flight each journey

6. **Crew**: information about the pilot, inflight crew- hostesses and stewards with their names, IDs, address, city, zip code, email address, etc.

7. **Class of service**: class code, class name. Also, it has status information, whether the passenger flown in the class he/she purchased or was upgraded.

8. **Booking channel**: booking reference, mode of booking (online, through app, at airport), promotion details--coupon codes, deals and discount used while booking

9. **Passenger profile information:** this is for each customer who is currently a member/non-member, his home airport, total miles traveled by the customer, and frequent flyer number

10.  **Payment**: modes of payment- online, card swipe, cash payment, or bank check. It also stores credit and debit card information

11.  **Partner/ vendors:** information about third-party partners and vendors needed such as fueling team, maintenance team, inflight food and snacks providers. The details would be vendor name, vendor ID, category, and type of service.

12.  **Employees**: employee’s name, employee ID, email address, phone number, address, city , zip code

|  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Business Processes** | **Conformed Dimensions** | | | | | | | | | |  |  |
| Airport | Aircraft | Passenger | Date | Time | Crew | Class of service | Booking channel | Passenger profile Information | Payment | Partner/  vendors | Employees |
| Issued Tickets | X | X | X | X | X |  | X | X | X | X |  | X |
| Reservations | X | X | X | X | X |  | X | X | X | X |  | X |
| Check-in/Bag Drop | X | X | X | X | X |  | X |  | X |  |  | X |
| Flight Activity | X | X | X | X | X | X | X | X | X |  |  |  |
| Customer Care  Interactions | X |  | X | X | X |  |  |  | X |  |  | X |
| Baggage  Tracking | X | X | X | X | X |  |  |  |  |  |  | X |
| Flight/Crew scheduling | X | X |  | X | X | X | X |  |  |  | X |  |
| Promotion tracking | X | X | X | X | X |  | X | X | X | X |  |  |
| Aircraft maintenance | X | X |  | X | X | X |  |  |  |  | X |  |
| Frequent flyer account credits | X | X | X | X | X |  |  |  | X | X |  | X |
| Onboarding services | X | X | X | X | X | X | X |  |  |  |  | X |
| Hotel/car rental | X | X | X | X | X |  | X |  | X | X |  | X |
| Training | X | X |  | X | X | X | X |  |  |  | X | X |

**Business processes:** The description of all the business processes is mentioned below

****

**d.      Calculations showing the size of Data Warehouse**

**i.      Make Assumptions on Number of Records for dimension tables and Number of Records per quarter on fact tables**

The below tables provide the size estimates for the dimensions and facts in the Datawarehouse for this company

**Dimension Tables:**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Table Name** | **Model Attributes** | **No. of attributes and Datatype (assumption)** | **Size of each row(in bytes)** | **Number of records** | **Size estimates of each table(in Bytes)** |
| **DimAirport** | Fields: Country, City, Name etc | 15 Fields  (10 varchar(20) + 5 int) | 240 | 50 | 12,000 |
| **DimAircraft** | Fields: Aircraft no. , Model, Type etc | 10 Fields  (7 varchar(20) + 3 int) | 176 | 56 | 9,856 |
| **DimPassengers** | Fields: Name ,Address, etc | 12 Fields  (10 varchar(20) + 2 int) | 228 | 50,000 | 11,400,000 |
| **DimTime** | Fields: Hours, Minute, Second | 8 Fields  (6 int + 2 Timestamp) | 30 | 86400 | 2,592,000 |
| **DimDate** | Fields: Day, Month, Year, Fiscalyear etc | 15 Fields  (8 varchar(20) + 7 int) | 204 | 3650  (past 10 years) | 744,600 |
| **DimEmployees** | Fields: EmployeeName, Address, Location etc | 10 Fields  (8 varchar(20) + 2 int) | 184 | 10,000 | 1,840,000 |
| **DimCrew** | Fields: CrewName, Department,Email,Salary etc | 15 Fields  (10 varchar(20) + 5 int) | 240 | 1000 | 240,000 |
| **DimServiceClass** | Fields: Classname, ClassCode, Status etc | 8 Fields  (6 varchar(20) + 2 int) | 144 | 10 | 1,440 |
| **DimBooking**  **Channel** | Fields: Mode of Booking, Promotion, RefNo. etc | 10 Fields  (8 varchar(20) + 2 int) | 184 | 100 | 18,400 |
| **DimPassengerProfile** | Fields: TotalMiles, Member(values- Yes or No), HomeAirport etc | 10 Fields  (8 varchar(20) + 2 int) | 184 | 50,000 | 9,200,000 |
| **DimPayment** | Fields: PaymentID, Mode of Payment, Credit Card No. etc | 15 Fields  (8 varchar(20) + 2 numeric+ 5 int) | 206 | 50,000 | 10,300,000 |
| **DimVendor** | Fields: VendorID, VendorName, Service etc | 8 Fields  (6 varchar(20) + 2 int) | 144 | 500 | 72,000 |

**Sum:** **36,430.296 KB**

**Size Estimates for Fact Tables:**

From the bus matrix we see that there a total of 12 business processes. Out of 12 business processes, some processes may have no facts and some may have many based on the business requirements. On average, let’s consider 2 Fact tables for each business process which makes total number of facts in this case as 24. As different Fact tables may be of different sizes, let’s assume 10 out of 24 fact tables to be of medium size, 10 as small size and 4 as large fact tables.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **Fact Tables** | **No. of Fact Tables** | **No. of attributes and Datatype (assumption)** | **Size of each row(in bytes)** | **Number of records per Quarter** | **Size Estimates** |
| **Small** | 10 | 30 Fields  (20 numeric+ 10 varchar(20)) | 320 | 90,000 | 28,800,000 |
| **Medium** | 10 | 30 Fields  (20 numeric+ 10 varchar(20)) | 320 | 900,000 | 288,000,000 |
| **Large** | 4 | 30 Fields  (20 numeric+ 10 varchar(20)) | 320 | 9,000,000 | 2,880,000,000 |

                                                                                                                                                        Total: 3.16 GB

Total size of Datawarehouse will be size of dimensions + size of fact tables

=0.036 GB + 3.16 GB

**Size of Datawarehouse=3.2 GB**

**e.      Analysis of relational vs. non-relational storage options**

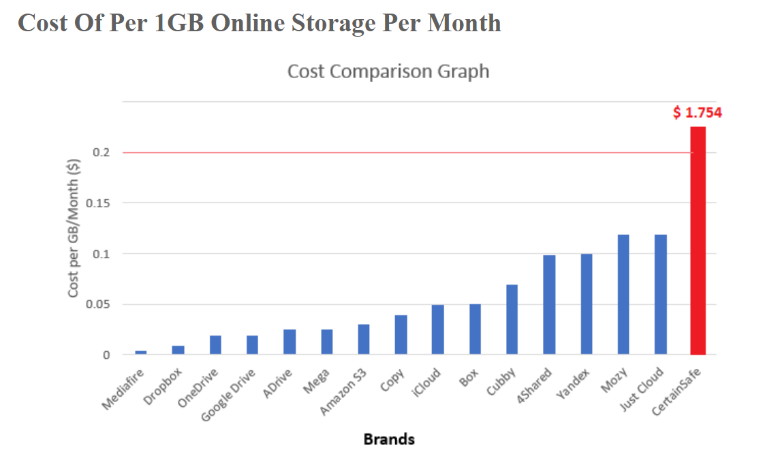
**i.      For each data source identified in (b) discuss the best way to store the data (relational vs. non-relational) and the underlying reasoning.**

**ii.      Present 2 options for big data storage and cost estimates.**

Some of the Sources identified for transactional and Bigdata are given below:

|  |  |
| --- | --- |
| **Data Sources(Transactional/ BigData)** | **Relational or Non- Relational Storage** |
| Human Resource Management system | · The data from HRM system is mostly transactional data. It involves activities corresponding to employee management processes of the business.  ·  As the data corresponding to this process is not huge, it’s feasible to store them in relational DB’s. |
| Ticket Reservation | · This process outputs both transactional as well as big data.  · As it involves click-stream data and other big data sources, it’s reasonable to store such huge datasets in non-relational storages. |
| Baggage tracking | · This process involves both transactional as well as big data sources.  · However, since big data collected in this case is not of much importance for analytical purposes, storing important information in relational DB is inexpensive. |
| Promotion Tracking | This process involves both transactional as well as big data sources.  ·However, since big data collected in this case is not of much importance for analytical purposes, storing important information in relational DB is inexpensive. |
| Flight Activity | Same as above |
| Aircraft Maintenance | Same as above |
| Flight/Crew Scheduling | · This process involves both transactional as well as Big Data sources.  ·  As it involves big data sources and as useful insights can be acquired from this data, it is reasonable to store this data using non-relational storages. |
| Flight tracking | Same as above |
| Third Party Sources | ·  This is the big data collected by separate vendors which can used in this industry.  ·  Since it’s big data, storing it in non-relational storage seems more accurate. |

**Big Data Storage options:**

****

|  |  |
| --- | --- |
| **Big Data Storage** | **Cost Estimate** |
| **Amazon AWS** | **Over 500 TB / month-  $0.0355 per GB** |
| **iCloud** | **1TB- $9.99 per month** |

**f.        Data Lake Architecture that describes how/where Data Warehouse and Big Data are stored and how they can be integrated for decision support. The architecture should include storage, processing and reporting layers for different types of data.**

* A data lake is a centralized repository that stores all structured and unstructured data at any scale. Data can be stored as-is. Different types of analytics—from dashboards and visualizations to big data processing, real-time analytics, and machine learning can be utilized to guide better decisions.
* Apart from transferring passengers from one location to another, the company also transfers thousands of terabytes of data. For handling these type of data, big data plays an important part in strengthening customer value and loyalty.  Following are some methods in which Compass Airlines can use big data.
* All business processes can be divided into RDBMS data and big data and will contain following activities.
* **RDBMS –**

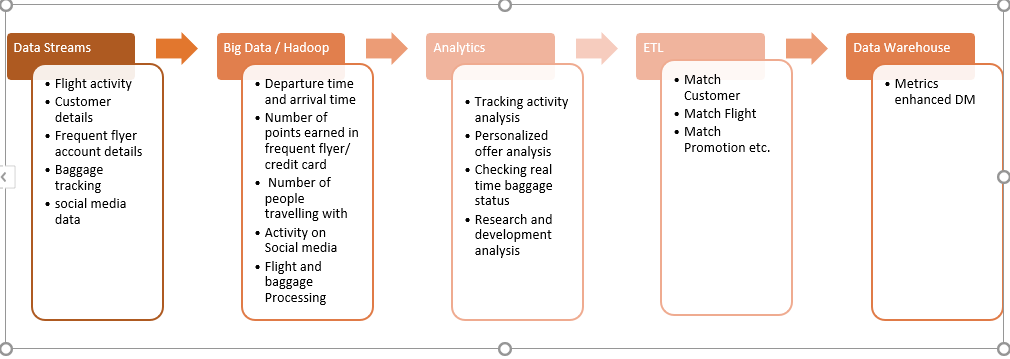
Crew: Crew information who is on-board

Aircraft Maintenance: Fuel requirement, maintenance requirement

Rental car or hotel booking: Hotel accommodation, car rentals and cab transports

Pilot training: Inflight or flight training

* **Big Data –**

1. Scheduling: Flight scheduling through website or on the spot, scheduled time, flight number, check in details, issue ticket, reservations
2. Flight activity: weather on arrival and while on-board, check-in, flight timing, Destination place, arrival and departure timings
3. Baggage delay tracking: checked in luggage, Realtime status of the language
4. Promotion tracking: Activities on social media, marketing and promotion activities
5. Customer care interactions: Interactions with the customer will be saved here for analysis
6. Frequent flyer account credits: number of points earned by the customer
7. Onboarding services: In flight food preferences, shopping, in the flight and at the airport

**Use cases for Data lake:**

Below are some use cases for big data. Use cases like this can be used for taking important decision

1. **Big data (Ingestion of semi-structured and unstructured data sources):** the company can have lots of data related to flight activity, customer data, social media activities etc. All the operational data such as number of flights booked, activity on social media, and number of points earned in frequent flyer facility demand new technology for research and development can be stored in data lake. This data not only helps customers to book their flight but to know their baggage real time activity, flight activity like how much time is remaining to reach the destination, etc. which has implications on many companies like travel insurance companies. This data can be unstructured, structured, text, image and can be in terabytes will require big data to process this data. A data lake is a great solution for storing all this data which is generally more difficult to store and can support near real-time analysis.
2. **Analysis and historical data storage:** When we store all the data mentioned above using big data, we can then use that data for further analysis. Sometimes data is used infrequently but does need to be available for analysis. A data lake strategy can be very valuable to support an active archive strategy.
3. **Data warehouse:** A data lake may contain data that isn't easily stored in a data warehouse or isn't queried frequently. The data lake might be accessed via federated queries which make its separation from the DW transparent to end users via a data virtualization layer.
4. **Storage of all organizational data to support reporting & analysis activities:** Some organizations wish to achieve a single storage repository for all types of data. Frequently, the goal is to store as much data as possible to support any type of analysis that might yield valuable findings.

**g**.       **Sample Reports or Schema that show how Big Data and Data Warehouse can be integrated to solve specific business decision problems relevant to the company/industry**

|  |  |
| --- | --- |
| **REPORT 1** | |
| **Report Name** | **Predicting highly personalized offers** |
| **Report Description** | ·        Airline booking has so many parts like frequent flyer account details, checked baggage, destination time and place, departure time and place, etc.  ·        Compass airline can collect this data to analyze highly personalized offer for customers. Big data solution can help airline to understand their customer individually, their preferences and behavioral pattern.  ·        Everything from the past purchases can be used to analyze flyer’s likely actions and can come up with custom made offer. This will increase sales of airline.  ·        Also, using this data we can analyze from what part of the state or region customers book a smaller number of flights and generate promotion offers to attract customers, which will help in boosting revenue of an airline. |

|  |  |
| --- | --- |
| **REPORT 2** | |
| **Report Name** | **Generate more accurate calls** |
| **Report Description** | ·        With every call from customer, compass airline can save lots of data from which customer service representative can understand nuances of every recorded customer interaction.  ·        Compass airline can analyze data from social media, to understand and gather more information from customers in real time.  ·        This will help Compass airline to deal with customer problems and gain customer loyalty. |

|  |  |
| --- | --- |
| **REPORT 3** | |
| **Report Name** | **Finding more safer flight route** |
| **Report Description** | ·        During air travel lots of data is generated per hour. If we consider average 6-hours flight time, 240 terabytes of data can be generated.  ·        This data can be used to analyze and mine using different technologies like machine- learning algorithms, data mining software to find out what can be potential safety issues. This analysis can be useful for avoiding any incident or accidents. |

|  |  |
| --- | --- |
| **REPORT 4** | |
| **Report Name** | **Checking real time baggage status** |
| **Report Description** | ·        All the baggage data can be collected and used to analyze and track the status of customers baggage.  ·        It is very important for customer that their baggage is safe and is with them in the same flight. Research and development team can work on this to build some application which will track and tells customers where exactly their bag is and send some notification, then customer will be happier and more loyal to the airline. |

**h.**     **Data Virtualization: What is data virtualization? How can the company can benefit from data virtualization? How does it fit into the organization data infrastructure?**

Data virtualization is any approach enabling data retrieval and/or data manipulation without the need to know where data is located or formatted. That means, regardless of form and storage site, data is easily accessible for end users. Analyses can be made on the data obtained, and there is no need to copy data when virtualization is being utilized. This allows data virtualization to offer rapid access to real-time data without wasting much of the resources. As companies grow and the sheer amount of data generated from their daily operations increase exponentially, data virtualization is of significance because it aggregates data from many disparate sources and transforms them into useful format. This allows companies to have a better understanding of their data. Additionally, in this age where data mining, big data, and artificial intelligence are prevalent and crucial to a company’s growth, data virtualization can tremendously assist these techniques since it breaks down data into digestible form. Data virtualization creates a central place where access to all data source is possible, so this makes it easier to implement security measurements to protect and monitor data.

**References**

1. <https://aws.amazon.com/big-data/datalakes-and-analytics/what-is-a-data-lake/>
2. <https://www.promptcloud.com/blog/five-interesting-use-cases-of-big-data-analytics-in-airline-industry/>
3. <https://ergonotes.com/how-much-does-a-gigabyte-of-online-storage-cost/>
4. Nazeri, Z. (2003). Application of aviation safety data mining workbench at American Airlines. The MITRE Corporation.
5. Mark Revels, Hélène Nussbaumer. Data Mining and Data Warehousing in the Airline Industry.
6. Rajendra Akerkar (2014). Analytics on big aviation data. Turning data into insights.
7. Kimball, R., Ross, M., Becker, B., Mundy, J., & Thorn. (2009). Kimballs Data Warehouse Toolkit Classics: The Data Warehouse Toolkit, 2nd. John Wiley & Sons