DataForge

Last Update as of 12/09

Analysis of Amtrak’s Data

This proposal contains a detailed description and plan of the database to highlight the insights using Amtrak data.

Created By:

Pranav Praveen Nair

Emil George Mathew

Abhishek Bhosale

Henry Kangten

ABOUT

MISSION STATEMENT

The mission is to analyze Amtrak’s 2021-2023 data to provide insights and recommendations on budgeting and costs using on-time performance, and ridership for various stations across cities. To analyze performance, recommend suggestions and story tell using intuitive visualizations   and data-driven decisions.

MISSION OBJECTIVES

The objective of this project is to focus on the following main metrics:
Creating the relevant Database, consisting of the tables and views for the Entities(station, state, route, stationMetrics, stateMetrics).

- ➔ Develop an ER diagram and relational schema that comprehensively captures the specified details and relationships between the aforementioned entities.

- ➔ Create SQL queries to retrieve and present data in alignment with the project objective. These include the following:
  - ◆ From the Business Analyst’s point of view:
    - Analyze the data from 2021-2023 to derive insights about key performance metrics in the data like ridership, budgeting, and on-time performance.
    - Storytelling using clear visualizations and enlightening with insights.
    - Determine procurement spending distribution by state and category to assess impact and correlation with other performance parameters like ridership.
    - Track the trend of employment spends for Amtrak workers in various states, explain the spending trends and visualize state-wise.
    - Suggest recommendations for improvement of current services and increase scope of the project.

  - ◆ From an Amtrak Manager’s point of view:
    - To gain any actionable insights to improve and optimize resource utilization, increase ridership and understand challenges that may be a major contributor to performance issues.
    - Identify problems that affect key metrics which lead to customer satisfaction.
    - Analyze and review employment spending distribution to align workforce requirements with operational needs.

BACKGROUND

PROJECT GOALS

Address challenges and recommend suggestions by identifying and analysing key performance metrics in the Amtrak data.

END USERS

Executive officers, decision makers, directors, operations team, data analysts.

BUSINESS TERMS

1. **Ridership**: The number of passengers using Amtrak services, typically measured at different stations and routes.
2. **On-Time Performance (OTP)**: A metric indicating the percentage of trains arriving on time, critical for assessing service quality.
3. **Procurement**: Refers to Amtrak's spending on goods and services, which may vary by city or station.
4. **Employment spending**: The total amount of money spent by Amtrak on employment costs like salaries, benefits and other expenses which can be analyzed by state.

ATTRIBUTES

1. **State** : {stateCode, stateName}
   Each state is identified by a 2-letter stateCode.
2. **Station**: {stationCode, stationName}
   Each station is identified by a 3-letter stationCode.
3. **StationMetrics: {**stationCode, stationMetric, stationMetricYear,stationMetricValue}
   Only the metrics that can be measured on a station level, are captured in this table. The metrics currently considered are (Procurement, Employment & EmploymentWage).
   Note: If in the future, the management has a different mission and wants to have visibility on other metrics too, the data can be appended without changing the ERD. This improves usability.
4. **StateMetrics: {**stateCode, stateMetric, stateMetricYear,stateMetricValue}
   Only the metrics that can be measured on a state level are captured in this table. The metrics currently considered are Ridership.
   Note: If in the future, the management has a different mission and wants to have visibility on other metrics too, the data can be appended without changing the ERD. This improves usability.

5. **Routes**: {routeID, routeName, routeType}
   Each route is identified by a 3-digit random number; there are 3 route types : (Long Distance, State Supported, Northeast Corridor)
6. **OnTimePerformance**: {routeID, stateCode, otpYear, otpValue}
   Each state on a route has a on time performance value for the years it was operational.
7. **Budget**: {budgetYearID, budgetType, budgetPlanYear, budgetTotal}
   A type of budget is planned each year. Each combination of the year and type is identified by a 3-digit random number in budgetYearID. There are 3 budget types - (Design, Construction & Deployment). The total assigned budget (in dollars) is the budgetTotal column.
8. **AllocatedBudget**: {budgetYearID, stationCode, allocatedBudgetYear, allocatedBudget}
   The budget planned in a year, can be spent/allocated (allocatedBudget) in the future years (allocatedBudgetYear).

CARDINALITIES

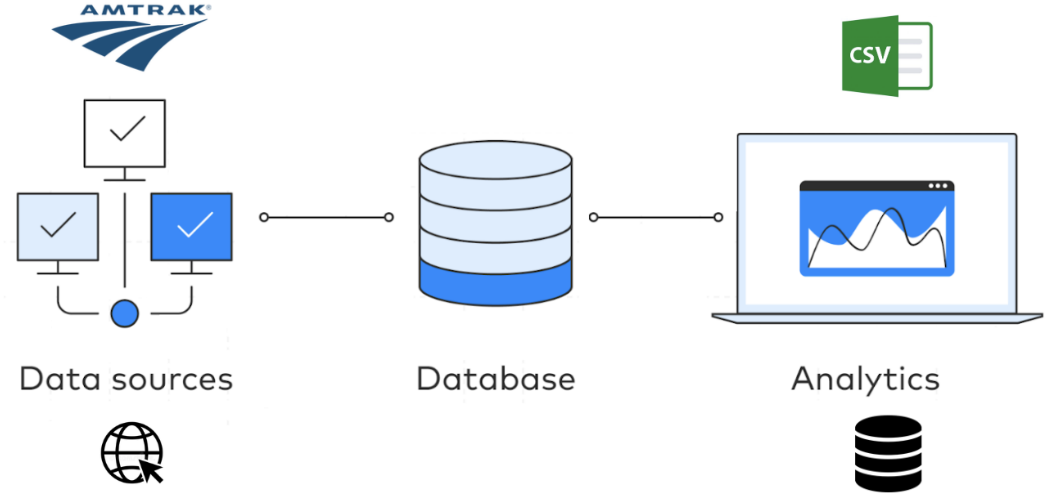
1. **Stations to State** : Many-to-One, each state can have multiple stations.
2. **Routes to Stations** : Many to Many, a route can pass through multiple stations; a station can be part of multiple routes.
3. **States to StateMetrics** : One to Many; each state can have multiple corresponding metrics.
4. **Stations to StationMetrics**: One-to-Many, each station can have multiple corresponding metrics.
5. **Budget to Station**: Many-Many, a station can have multiple types of budget planned for different years, a budget type can be decided for multiple stations.

DATA SOURCES & PIPELINES

Data Sources: The data has been downloaded from the Amtrak website and collected in an xlsx file.

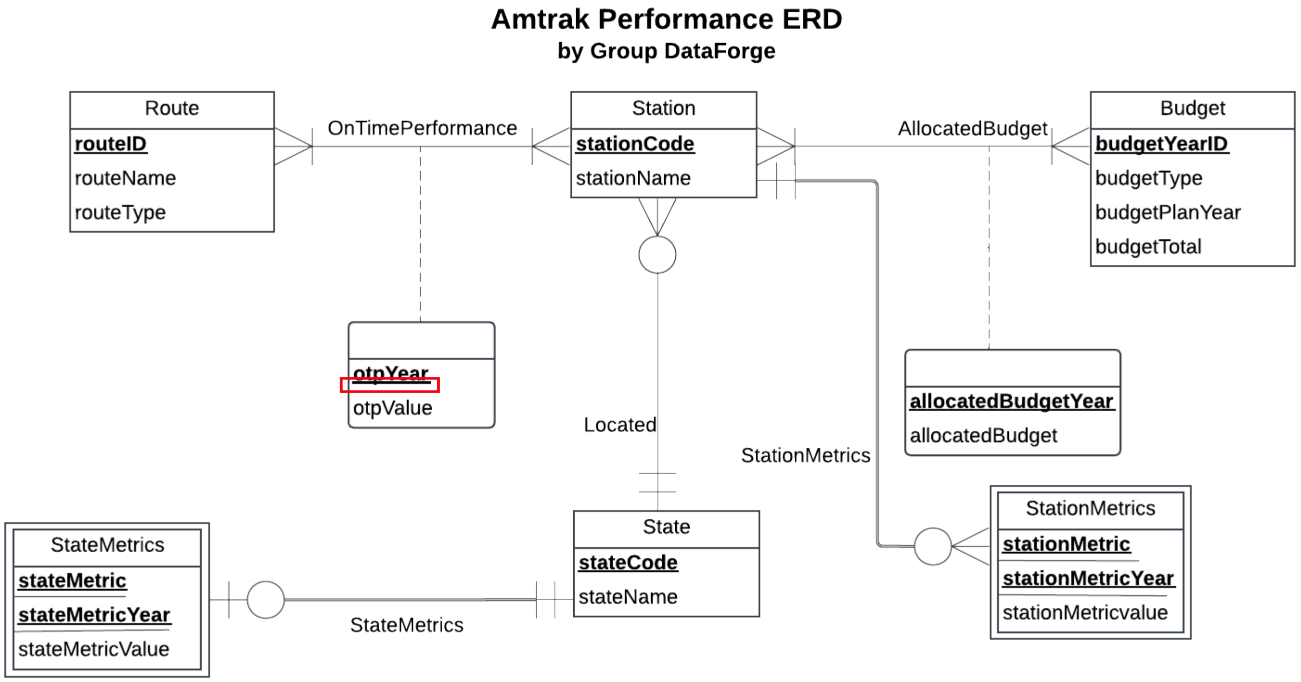
1. State Fact Sheets | Amtrak
2. Reports & Documents | Amtrak
3. Train Routes | Amtrak
4. Wikipedia | Amtrak Stations

Data Pipeline: The diagram shows the data pipeline of Amtrak Data.



DATABASE DESIGN

ENTITY RELATIONSHIP DIAGRAM


Amtrak Performance ERD
by Group DataForge

RELATIONAL SCHEMA

State (**stateCode**, stateName)

Station (**stationCode**, stationName, *stateCode*)

StateMetrics (***stateCode***,**stateMetric**,**stateMetricYear**,stateMetricValue)

StationMetrics (***stationCode***,**stationMetric**,**stationMetricYear**,stationMetricValue)

Route (**routeID**, routeName, routeType)

OnTimePerformance (***routeID***, ***stationCode***, **otpYear**, otpValue)

Budget (**budgetYearID**, budgetType, budgetPlanYear, budgetTotal)

AllocatedBudget (***stationCode***, ***budgetYearID***, **allocatedBudgetYear**, allocatedBudget)

BUSINESS RULES

R1: When a state is deleted or updated in the database (if Amtrak discontinues to serve a particular state), the corresponding stations also become inactive and are deleted/updated from the database.

R2: When a state is deleted or updated in the database (if Amtrak discontinues to serve a particular state), the corresponding metrics data is also deleted/updated.

R3: When a station is deleted or updated in the database (if Amtrak discontinues to serve a particular station), the corresponding metrics data is also deleted/updated.

R4: When a route is deleted or updated in the database (if Amtrak discontinues to serve a particular route), the corresponding on time performance data is also deleted/updated.

R5: When a station is deleted or updated in the database (if Amtrak discontinues to serve a particular station), the corresponding on time performance data is also deleted/updated.

R6: When a budget along with its allocation for the following years has been planned for a year, it cannot be deleted/updated in the database.

R7: When a budget for a station has been planned in a year, both budget & station cannot be deleted/updated in the database.

REFERENTIAL INTEGRITIES

Business Rule	Relation	Foreign Key	Base Relation	Primary Key	ON DELETE	ON UPDATE
R1	Station	stateCode	State	stateCode	CASCADE	CASCADE
R2	StateMetrics	stateCode	State	stateCode	CASCADE	CASCADE
R3	StationMetrics	stationCode	Station	stationCode	CASCADE	CASCADE
R4	OnTimePerformance	routeID	Route	routeID	CASCADE	CASCADE
R5	OnTimePerformance	stationCode	Station	stationCode	CASCADE	CASCADE
R6	AllocatedBudget	budgetYearID	Budget	budgetYearID	NO ACTION	NO ACTION
R7	AllocatedBudget	stationCode	Station	stationCode	NO ACTION	NO ACTION

SAMPLE DATA FOR EVERY RELATION

1. Station

   ('ATN', 'Anniston', 'AL'),

   ('BHM', 'Birmingham', 'AL'),

   ('TCL', 'Tuscaloosa', 'AL'),

2. StateMetrics

   ('AL', 'Employment', 'FY21', 13),

   ('AZ', 'Employment', 'FY21', 12),

   ('AR', 'Employment', 'FY21', 26),

3. StationMetrics

   ('ATN', 'Ridership', 'FY21', 1948),

   ('BHM', 'Ridership', 'FY21', 14935),

   ('TCL', 'Ridership', 'FY21', 3720),

4. OnTimePerformance

   (100, 'AL', 'FY21 ', 0.546),

   (101, 'AZ', 'FY21 ', 0.363),

   (102, 'AZ', 'FY21 ', 0.271),

5. AllocatedBudget

   ('137', 'Design', '2020', '275000')

   ('177', 'Construction''2020', '2638000')

   ('239', 'Design', '2023', '0')

These are the sample data for every relation.