Building an Azure Data Lake for Bike Share Data Analytics

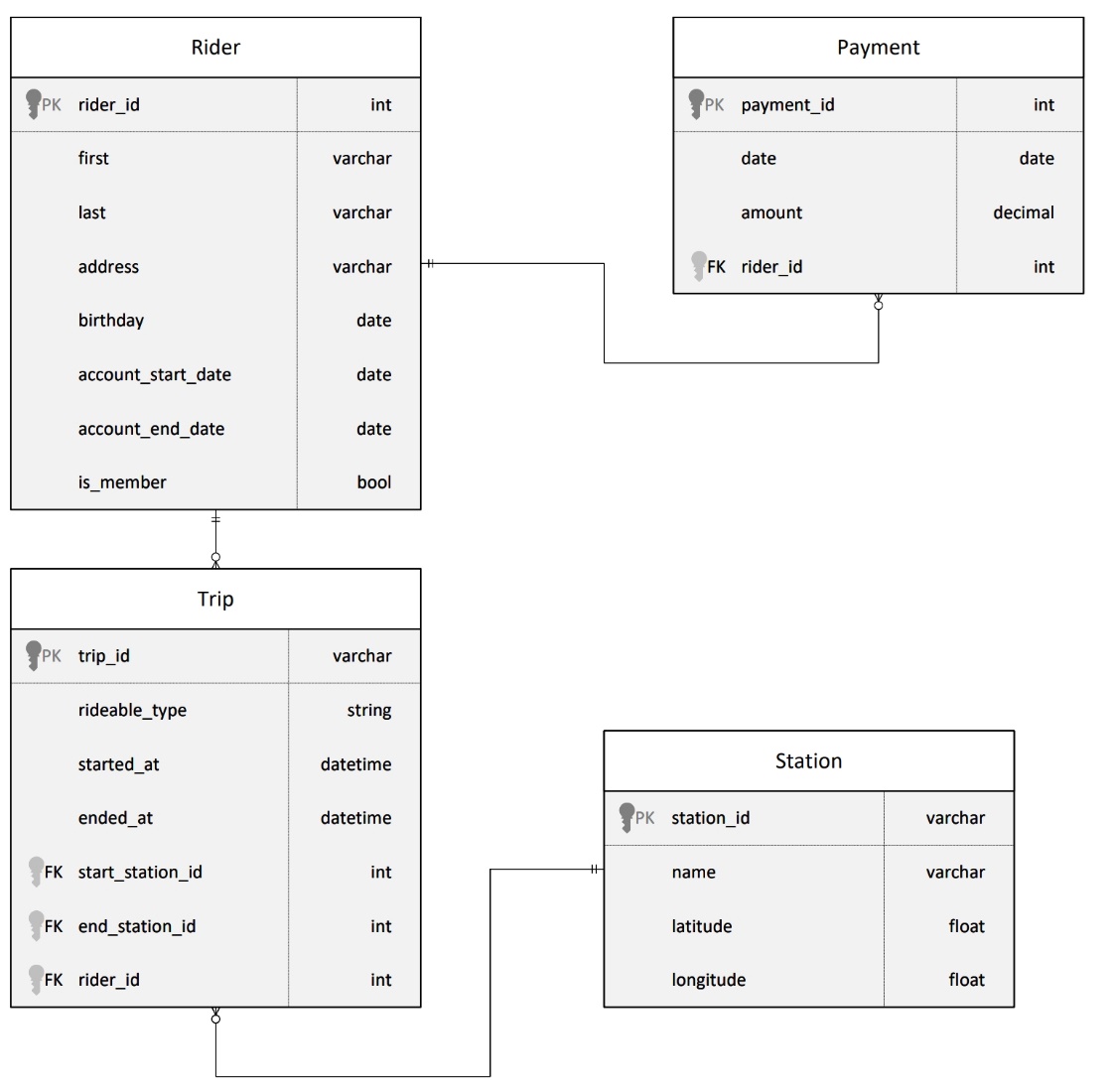
**Project Overview**

In this project, you'll build a data lake solution for Divvy bikeshare.

Divvy is a bike sharing program in Chicago, Illinois USA that allows riders to purchase a pass at a kiosk or use a mobile application to unlock a bike at stations around the city and use the bike for a specified amount of time. The bikes can be returned to the same station or to another station. The City of Chicago makes the anonymized bike trip data publicly available for projects like this where we can analyze the data.

Since the data from Divvy are anonymous, we have generated fake rider and account profiles along with fake payment data to go along with the data from Divvy. The dataset looks like this:

:



Relational ERD for the Divvy Bikeshare Dataset (with fake data tables)

The goal of this project is to develop a data lake solution using Azure Databricks using a lake house architecture. You will:

* Design a star schema based on the business outcomes listed below;
* Import the data into Azure Databricks using Delta Lake to create a Bronze data store;
* Create a gold data store in Delta Lake tables;
* Transform the data into the star schema for a Gold data store;

**The business outcomes you are designing for are as follows:**

1. Analyze how much time is spent per ride
   * Based on date and time factors such as day of week and time of day
   * Based on which station is the starting and / or ending station
   * Based on age of the rider at time of the ride
   * Based on whether the rider is a member or a casual rider
2. Analyze how much money is spent
   * Per month, quarter, year
   * Per member, based on the age of the rider at account start
3. EXTRA CREDIT - Analyze how much money is spent per member
   * Based on how many rides the rider averages per month
   * Based on how many minutes the rider spends on a bike per month

On the next page are instructions for logging into an Azure account where you can configure the resources, Azure Synapse Workspace, and data storage to complete the project.

**Project Environment**

In order to complete this project, you'll need to use these tools:

* Access to Microsoft Azure. Instructions for accessing an Azure account where you can create the resources necessary for the project are on the previous page.
* Hint: To view your DBFS files, enable the DBFS file browser in Databricks by going to Admin Console -> Workspace Settings -> Advanced
* Hint: If you are going to use PySpark Pandas, make sure you create your Spark Cluster using a Databricks runtime >= 10.0

**Starter Code and Data**

You'll need the [dataset](https://video.udacity-data.com/topher/2022/March/62420bb1_azure-data-lakehouse-projectdatafiles/azure-data-lakehouse-projectdatafiles.zip) from the course resources menu in the left navigation bar. Instructions for how to use the data in your project are on the next page in the project instructions

**Project Instructions**

The goal of this project is to develop a data lake solution using Azure Databricks using a lake house architecture. You will:

* Design a star schema based on the business outcomes below;
* Import the data into Azure Databricks using Delta Lake to create a Bronze data store;
* Create a gold data store in Delta Lake tables;
* Transform the data into the star schema for a Gold data store;

You'll implement these requirements by creating a Python notebook, or notebooks in the Azure Databricks workspace. Detailed requirements for what should be in the notebook can be found in the project rubric. You'll submit your notebook(s) and they'll be assessed against this [rubric](https://review.udacity.com/#!/rubrics/4817/view) when you submit your project.

**The business outcomes you are designing for:**

1. Analyze how much time is spent per ride
   * Based on date and time factors such as day of week and time of day
   * Based on which station is the starting and / or ending station
   * Based on age of the rider at time of the ride
   * Based on whether the rider is a member or a casual rider
2. Analyze how much money is spent
   * Per month, quarter, year
   * Per member, based on the age of the rider at account start
3. EXTRA CREDIT - Analyze how much money is spent per member
   * Based on how many rides the rider averages per month
   * Based on how many minutes the rider spends on a bike per month

Design a star schema for these business outcomes. **Create a PDF of the schema** you'll submit along with your Azure Databricks notebook file(s).

Graphical user interface, text, application

Description automatically generated

**Submission Instructions**

Here is a list of things to submit for this project. Create a zip file containing:

* PDF of the star schema you designed based on the relational diagram and the business problems outlined
* Your Azure Databricks notebook(s)