|  |  |
| --- | --- |
| **PROJECT SPECIFICATION** | |
| **Azure Data Lakehouse Project** | |
| **Star Schema Design** |  |
| CRITERIA | MEETS SPECIFICATIONS |
| The student will be able to generate fact tables based on a business need and a relational model | The star schema should have at least two fact tables. One should be related to trip facts and another should be related to payment facts. The trip fact should have a fields for trip duration and rider age at time of trip. The payment fact should have a field related to amount of payment. |
| The student will be able to generate dimension tables based on business needs and a relational model | The star schema should have dimensions related to the trip fact table that are related to: riders, stations, and dates. The schema should have dimensions related to the payment fact table that are related to: dates and riders. |
| **Extract Step** |  |
| CRITERIA | MEETS SPECIFICATIONS |
| Produce Spark code in Databricks using Jupyter Notebooks and Python scripts | The notebook should contain Python code to extract information from CSV files stored in Databricks and write it to the Delta file system. |
| Use distributed data storage using Azure Data Storage options | The notebook should contain Python code that picks files up from the Databricks file system storage and writes it out to Delta file locations. |
| **Load Step** |  |
| CRITERIA | MEETS SPECIFICATIONS |
| Implement key features of data lakes on Azure | The notebook should contain code that creates tables and loads data from Delta files. The learner should use spark.sql statements to create the tables and then load data from the files that were extracted in the Extract step. |
| **Transform Step** |  |
| CRITERIA | MEETS SPECIFICATIONS |
| Use Spark and Databricks to run ELT processes by creating fact tables | The fact table Python scripts should contain appropriate keys from the dimensions. In addition, the fact table scripts should appropriately generate the correct facts based on the diagrams provided in the first step. |
| Use Spark and Databricks to run ELT processes by creating dimension tables | The dimension Python scripts should match the schema diagram. Dimensions should generate appropriate keys and should not contain facts. |
| Produce Spark code in Databricks using Jupyter Notebooks and Python scripts | The transform scripts should at minimum adhere to the following: should write to delta; should use overwrite mode; save as a table in delta. |