Final Project Architecture Blueprint:

Section 1 – Business Problem

We are building a real-time NYC CitiBike trip prediction analysis pipeline that helps analyst track expected trips and demand using Open-Meteo's Forecast API. The system will surface demand trends over the next 7 days using forecast data to provide clear trip projections and show how different weather features are influencing customer behavior in NYC.

Section 2 – Data Sources

NYC CitiBike has a website where they post monthly and yearly downloadable CSVs with their data that includes the date and time of the trip along with their customer type (casual or member), station data, and ride specific data like start and end times. We plan to standardize the dates on these CSVs for the years 2022, 2023, and 2024, then call the Open-Meteo weather forecast API to retrieve historical daily forecast records from that same period. We will count the historical daily number of trips, member trips, and casual trips using the CSV data and standardize the dates on the historical records from the API, then create a curated table where we join both tables together to create a final table that holds trip information and forecast data like min/max temp, max wind speed, and precipitation to train our ML model. We will then utilize live streaming from the Open-Meteo API to feed fresh data into our ML model for predictions and then feed the scored predictions into our Looker Studio dashboard to provide real-time analytics.

Section 3 – Cloud Architecture

We'll have Cloud Scheduler call our Cloud Run service (open-meteo) which will then fetch the 7-day forecast for NYC. It will then build one JSON message per forecast date with included features like 'forecast_date,' max/min temperatures, precipitation, and more. Those messages will then be published to Pub/Sub and then serve as the input to Dataflow. Dataflow will stream "Exactly once," putting the raw forecast streaming data into a designated raw table. We'll use a SQL query to take the 7 most recently ingested rows (our fresh 7-day forecast) and insert them into our ML model to predict member and casual trip demand, then join the two predictions in a curated predictions table. Finally, this table will be used in Looker Studio to provide real-time analytics and insights.

Section 4 – ML Plan

We will use 2 regression models (one for member trips and one for casual trips) to predict the number of daily member and casual trips over the next 7 days using forecast data. We will use the features max/min temperature in Fahrenheit, precipitation in inches, wind

speed in MPH, and then engineer 3 new features. We will use “dow” which is the day of the week where 1 is Sunday and 7 is Saturday, “month” which is the number where the month falls into the full year where 1 is January and 12 is December, and we will create “is_weekend” where it’s 1 if the date is on a weekend and 0 if not. We will use ML.EVALUATE to evaluate model performance and try different models during initial testing to see what gives the best results. Our models will output the predicted number of daily member and casual trips, then we will add them to create daily total trips.

Section 5 – Dashboard KPIs

We will create 2 dashboards: 1 that displays trends in historical data and 1 that is a real-time analytics dashboard. The historical dashboard will examine trends like historical number of daily trips over time, daily trips vs. average temperature, and the total distribution of trips by day of the week.

Our real-time analytics dashboard will display metrics like the predicted total number of member trips, casual trips, and total trips over the next 7 days along with a card that shows the most recent ingestion timestamp in UTC. We will have 2 time-series bar graphs that display the number of predicted member and casual trips by day to see how demand varies throughout the week. We will also have 2 time-series line graphs where 1 will display the number of total predicted trips vs. average temperature by day and the other will display the number of total predicted trips vs. forecast max wind speed by day.

Both dashboards will serve to both provide insights about CitiBike consumer trends over the past 3 years while also giving up-to-date predictions of consumer behavior based on weather forecasts, allowing analysts to examine how weather influences consumer behavior in real time.

Section 6 – Risk and Mitigations

We will initially start with 2022-2024 data as a proof of concept and to provide reliable predictions based on recent years where there aren’t large anomalies like covid at play. We will try to incorporate more years into our training data and model, but if we notice that it significantly hurts model performance then we will likely return to the 2022-2024. CitiBike has also changed the way they format their downloadable CSVs throughout the years which may cause issues in both formatting for table joining and ML training. We will try our best to normalize and standardize the data, but we will likely revert to the original time frame again if we notice poor performance as a result of incorporating the new data.