Final Project Architecture Workshop (Blueprint)

Date generated: 2025-08-21

Use this as your *structured design doc*. Replace all placeholders with your team’s plan.

Section 1 — Business Problem (1 paragraph)

Hotels face significant uncertainty in managing cancellations, occupancy, and staffing because booking behavior can change quickly as conditions evolve closer to arrival dates. While historical booking data provides strong baseline patterns, it does not account for real time external context such as weather conditions. This project addresses the problem of understanding and forecasting hotel booking cancellations by combining historical hotel booking data with both historical and real time weather data. The goal is to help hotel operators anticipate cancellation risk under current weather conditions and make more informed operational decisions related to staffing, inventory, and pricing.

Section 2 — Data Sources (batch + streaming) with access details

Batch Data Source

The batch dataset consists of historical hotel booking records enriched with historical daily weather data. The hotel booking data includes reservation attributes such as lead time, stay length, customer type, and cancellation status. Historical weather variables such as temperature, precipitation, wind speed, and weather codes are joined to bookings based on hotel location and arrival date. This curated dataset is stored in BigQuery and serves as the primary training data for the BigQuery ML cancellation prediction model.

Streaming Data Source

The streaming data source is a real time weather API that provides current weather observations for the hotel location. Weather data is fetched approximately every ten minutes using a scheduled cloud service and ingested into a BigQuery streaming table. Each record includes timestamped measurements such as temperature, precipitation, wind speed, and weather conditions. This data is used to provide real time context for dashboards and model interpretation.

Section 3 — Cloud Architecture (diagram + narrative)

ASCII draft: [API] -> [Cloud Function producer] -> [Pub/Sub topic] -> [Dataflow template] -> [BigQuery] -> [Looker]

Architecture Narrative

The pipeline begins with a scheduled weather ingestion service that queries a public weather API at regular intervals. The raw weather observations are published to a Pub/Sub topic and processed by a Dataflow streaming job. Dataflow performs light transformations and writes structured rows into a BigQuery streaming table. Historical hotel booking data and historical weather data are stored in BigQuery batch tables. BigQuery ML is used to train a classification model that predicts the probability of booking cancellation. Looker Studio connects directly to BigQuery to visualize historical trends, real time weather conditions, and model driven insights.

Logical Flow

Weather API → Cloud Scheduler → Pub/Sub → Dataflow → BigQuery (streaming) Historical Hotel Data + Historical Weather → BigQuery (batch) → BigQuery ML BigQuery → Looker Studio Dashboard

Section 4 — ML Plan (BQML): target, features, metric, scoring mode

Prediction Target

The model predicts whether a hotel booking will be canceled. This is a binary classification problem with the target variable is_canceled.

Features

Features are drawn from historical booking behavior and historical weather conditions and include:

Lead time

Length of stay

Customer type

Market segment

Deposit type

Booking changes

Repeated guest indicator

Temperature metrics

Precipitation

Wind speed

Model Type

A logistic regression classification model is trained using BigQuery ML.

Metric

Model performance is evaluated using standard classification metrics including accuracy, precision, recall, and AUC.

Scoring Mode

The model is trained on batch data and used for batch predictions. Real time weather data is used as contextual information for interpretation and dashboard insights rather than direct causal scoring.

Section 5 — Dashboard KPIs (3–5) with definitions and SQL sources

The Looker Studio dashboard visualizes both real time weather conditions and historical cancellation behavior to provide operational insight into cancellation risk drivers.

Weather Time Series

This chart displays real time temperature and wind speed over time using the streaming weather data pipeline. It provides situational awareness of current environmental conditions that may contextualize booking behavior. Source: BigQuery streaming weather table.

Precipitation Time Series

This chart shows real time precipitation levels over time, highlighting periods of rainfall intensity. Although precipitation occurs intermittently, spikes provide useful context when evaluating short term booking behavior and cancellations. Source: BigQuery streaming weather table.

Cancellation Rate by Deposit Type

This bar chart compares cancellation rates across different deposit policies. It shows that refundable bookings have a higher cancellation rate than non refundable or no deposit bookings, reinforcing the importance of deposit structure in cancellation risk management. Source: Historical hotel booking data in BigQuery.

Cancellation Rate by Season

This visualization shows how cancellation rates vary by season, with higher rates observed in fall and spring relative to winter. This highlights seasonal patterns in traveler behavior that are independent of short term weather conditions. Source: Historical hotel booking data with derived seasonal features.

Cancellations by Lead Time

This chart illustrates how cancellations increase as lead time increases. Bookings made further in advance exhibit substantially higher cancellation rates, confirming lead time as one of the strongest predictors of cancellation behavior. Source: Historical hotel booking data in BigQuery.

Cancellation Rate by Temperature

This chart groups historical bookings into temperature bins based on arrival date weather and shows the associated cancellation rate. While variation exists across temperature ranges, the relationship is weaker than core booking attributes, reinforcing that weather acts as contextual rather than causal input. Source: Historical hotel bookings joined with historical weather data in BigQuery.

Section 6 — Risks & Mitigations (Devil’s Advocate) + Prompt

Prompt: What is the most significant risk to the success of this architecture?

Response: The most significant risk to the success of this architecture is overinterpreting real time weather signals as direct drivers of hotel booking cancellations. The historical data shows that core booking attributes such as lead time, deposit type, and customer history are far stronger predictors of cancellation behavior than short term weather conditions. If weather data is treated as causal rather than contextual, stakeholders could draw incorrect conclusions or make suboptimal operational decisions. To mitigate this risk, the system is designed so that weather data provides real time context rather than standalone decision signals. The machine learning model prioritizes booking attributes, while weather features are included to enrich interpretation. The dashboard reinforces this framing by presenting weather time series alongside cancellation trends rather than implying direct causality.

Section 7 — Milestones & Ownership (30/60/90 or weekly)

Week 1 Loaded and explored historical hotel booking data in BigQuery and validated key cancellation drivers.

Week 2 Integrated historical weather data and created a curated batch table aligned on arrival date and location.

Week 3 Built the real time weather ingestion pipeline using scheduled API calls and BigQuery streaming tables.

Week 4 Trained and evaluated a BigQuery ML logistic regression model to predict booking cancellations.

Week 5 Designed and finalized the Looker Studio dashboard, connected real time and batch data sources, and prepared demo and documentation materials.