*Please refer to the Employment Application and NDA you have signed prior to reaching this stage in process. We will be monitoring to ensure no one is sharing the case study online and/or asking for assistance.*

# Customer Engineer On-Site Case Study & Presentation Interview Guidelines

Your on-site presentation interview has two parts: a GCP Case Study presentation and a "Candidate's Choice" presentation. You'll be delivering both of these presentations in one hour (60 minutes). Please organize this hour as you see fit; we recommend that you plan to spend roughly ⅔ (40 minutes) of your time on the GCP Case Study. Your audience will be chiming in with questions throughout your presentations, so budget your time and plan your content knowing that Q&A will be interspersed throughout the hour.

It's a good idea to use Google Slides in lieu of Powerpoint or Prezi. You may find the GCP Solution Icons for architectural diagrams helpful in creating your Case Study slides.

**Watch this video from the CE team** regarding the presentation; it provides helpful advice and tips to keep in mind when preparing your presentation.

1) **GCP Case Study.** Present a recommendation and demo for a GCP-based Solution to **one** of the scenarios listed below. A more detailed overview of each scenario is found later in this document.

   *Scenario 1: IoT: Aggregating and Analyzing Fitness Data*
   *Scenario 2: Real-Time Data Ingest: Traffic Telemetry*
   *Scenario 3: Web Serving: Autocomplete*
   *Scenario 4: Big Data: Aggregating and Analyzing Doubleclick Data*

2) **Candidate's Choice.** Educate your audience on something you're passionate about outside of work - the topic is up to you. Maybe you want to teach us how to DJ, walk us through how you create a killer homemade salsa, or convince us that Lord of the Rings is a better trilogy than the original Star Wars. Feel free to be creative!

Your presentation audience will include some customer engineers and some sales representatives. Be prepared to handle a meeting that has video participants.

## Part 1. GCP Case Study: Deliver a recommendation & demo for a GCP-based Solution

In this presentation, your audience will be acting as a customer group and you should act as though you are a Google Customer Engineer. Select **one** of the scenarios and prepare at least one recommended architecture on Google Cloud Platform. **The goal of this portion of this interview is to assess your ability to quickly acquire technical knowledge and deliver a credible presentation in a customer setting (which is potentially stressful).**

### General Guidelines and Advice

- Limit the presentation to 5-7 slides
- Have an architecture that you believe solves the problem; be prepared to defend your choices and discuss tradeoffs.
- **The Googlers you're presenting to work on these products, but are role-playing as though they're customers who don't.** Understand that the audience definitely knows more about the product than you do. Expect questions that will push you beyond your knowledge boundaries.
- **To that note, remember that this is a simulated customer meeting. Act accordingly.**
- **Have a working demo.**
- Read through the Q&A at the end of this document to help you prepare for the presentation.

## ***Scenario 1. IoT: Aggregating and Analyzing Fitness Data***

A National Health Club company (customer) is looking to modernize their member profile system by empowering their members to track their individual key health and wellness metrics on a daily basis using their mobile device. The customer would like to allow members to choose any mobile app and/or wearable device to collect the data so as long as the data can be exported or accessed through an API request to store in a centralized repository.

For the use case, current supported platforms are Fitbit, Google Fit, and Apple Health app. The customer expects that this system will: 1) improve the accuracy of metrics collected for their Personal Training department in between training sessions and 2) increase member retention by sharing data analysis with members and offering targeted incentive programs based on member's progress and results.

The fitness data may come in a variety of file formats (XML, JSON, CSV, etc) for each member and could be uploaded into cloud storage daily. The customer intends to provide access to each member to the API link or to upload the data. The customer anticipates about 50% of their members will participate which is roughly 1,000,000 members and each data file is no larger than 100Mb. The customer requires that one year's worth of data for each member be available in the system coinciding with the member's membership start date. Once member's fitness data has been processed by the system, there is no need to retain the export files.

The customer is looking for some recommendations on how to set up a system that allows them to collect the data and analyze trends and patterns over time and across members through a reporting dashboard. Initially, the 3 key areas they would like to analyze is sleep (hours per day), calories consumed (per day), and calories burned (per day) eventually adding in additional metrics around BMI, heart rate monitoring, etc. The individual member's metrics should be analyzed and compared with national health recommendations based on the individual's age, weight, height, etc to get a baseline of how the member is doing compared against the national health recommendations published by the Center for Disease Control and Prevention (CDC).

**Minimum Expectations**

- One recommended architecture and prototype running on GCP with details on how data flows through the system and descriptions of each component
- Use sample health datasets
- Build a data pipeline
- Create visualization for fitness trends and member wellness
- At least one alternative architecture with pros/cons
- Estimated pricing for each architecture

**Some Considerations**

- Incorporating open source technology is preferred
- How can you accommodate for the variety of data sources to be ingested?
- Have fun and be healthy - use your own data metrics.
- What type of visualization reporting mechanisms are supported?

## ***Scenario 2. Real-Time Data Ingest: Traffic Telemetry***

A major US city has installed a traffic tracking system that can report on the speed of vehicles passing through a given street at a given time period, thereby estimating the flow of traffic at a given time. This ever-changing data set is recorded at periodic intervals and made available by public API. Currently, about 150,000 readings are emitted every day citywide, which continues to grow as increasingly inexpensive sensors are installed by the city. The city has requested a recommendation of possible solutions that will allow this collection and analysis of the API data; the primary focus is on the underlying data architecture, a secondary focus is on organizing and analyzing this data. Devise a method to ingest this data as it arrives, perform any necessary transformations to make it useful, and store the result in a location where it can be analyzed at scale in a timely fashion. (The city will also consider any recommendations for visualization or other end-user solutions that can help organize this information, and make it accessible and useful.)

**Minimum Expectations**

- One recommended architecture and prototype running on GCP with details on how data flows through the system and descriptions of each component
- Use the City of Chicago traffic dataset, which contains both historical data and the current feed. Gathering live data through the linked API is recommended.
- Capture multiple data points for a sustained period of time, sufficient to show how your architecture can serve for both ongoing data ingest and historical data analysis.
- At least one alternative architecture with pros/cons.
- Estimated pricing for each architecture.

**Some Considerations**

- Incorporating open source and cloud-native technology is preferred.
- Real world traffic data is frequently updated in real time; approach the problem with this in mind.
- What are some ways the data could be captured and transformed on the fly?
- What types of systems could be used to store and analyze the telemetry data?
- What methods could a customer use to detect interesting / unusual events in the data stream?
- What would happen if we scaled this system to 10 cities? To 100? To every city in the world?

## ***Scenario 3. Web Serving: Autocomplete***

A e-commerce customer wants some advice on how to implement an autocomplete feature that provides real-time, low latency suggestions as a user types in a search phrase for their global user base. They currently have a product catalogue of around 3 million objects and only want to autocomplete the product name.

### Minimum Expectations

- One recommended architecture and prototype running on GCP with details on how data flows through the system and descriptions of each component
- Use a real dataset from Best Buy
- At least one alternative architecture with pros/cons
- Estimated pricing for each architecture

### Some Considerations

- Incorporating open source technology is preferred
- Serving a global user base with low latency
- What are some ways they could generate the autocomplete data?
- What types of systems could be used to store and serve the autocomplete data?

## ***Scenario 4. Big Data: Aggregating and Analyzing Doubleclick Data***

An advertising agency (customer) manages marketing for many different companies (clients). They purchase advertising on a variety of platforms, including Doubleclick. This resulting impression data is provided as a daily dump for each of their clients, typically 25GB per client per day. They have over 500 clients, amounting to just over 12 TB of data generated every day.

The Doubleclick data is in a star schema and a series of CSV files for each client are deposited into cloud storage every night. Since it's a distributed system, not all data is 100% accurate and may be changed or added in subsequent daily dumps, e.g., Friday's dump may include a lot of impressions that actually happened on Monday.

The customer (advertising agency) is looking for some recommendations on how to set up a system that allows them to analyze trends and patterns over time and across clients. They also need to provide reporting and/or exports to individual clients on only their data.

### Minimum Expectations

- One recommended architecture and prototype running on GCP with details on how data flows through the system and descriptions of each component
- At least one alternative architecture with pros/cons
- Estimated pricing for each architecture

### Some Considerations

- Incorporating open source technology is preferred
- How fast would the processing take?
- How could they extend the system to include data from other ad exchanges?
- What type of reporting mechanisms are supported?

## Part 2. Candidate’s Choice: Educate your audience on a topic you’re passionate about

In this portion of the panel, your audience will be themselves. Select a topic that you’re passionate about (ideally one outside of work - i.e., a hobby, interest, or volunteering activity) and share your knowledge with us. **The goal of this portion of the interview is to assess your ability to “own the room” and evaluate how well you’re able to communicate and educate an audience on a topic once you acquire a deep expertise. We want to see how you present on a topic that you’re the expert on, and the Googlers are not.**

### General Guidelines and Advice

- You can use any presentation means/method you choose
- Have fun with it!
- Choose a topic you’re passionate about. Some teach us a new skill (e.g., tap dancing!). Some tell us about their volunteer or charitable involvements. If you’re not comfortable presenting on something personal, some candidates pitch publicly available information about their current product.

## Advice from Recently Hired Customer Engineers

- Ask a friend or family member if he or she will observe you giving a **dry run of your presentation** the week before you’re scheduled to interview to act as your test audience and provide honest feedback.
- Take the Toastmasters Crash Course.
- Record yourself giving the presentation.
- Remember this is a **sales** demo. State the business outcome/goal first.
- Ensure your audience understands and is on the same page with you before you move to your next point.
- End with cost estimates for solution and qualify next steps.
- **Watch this video from the CE team** regarding the presentation; it provides helpful advice and tips to keep in mind when preparing your presentation.

## FAQs: Preparing for your CE Presentations

**Use this Q&A to guide your preparations for your onsite**

### General FAQs

**Question**: Who is our audience?
**Answer:** Mix of business and IT. Really it will be a couple reps and some fellow SEs and none of them will take the roles extremely seriously. But you will probably get a question or two on cost/value.

**Question**: How granular are we expect to be in our responses?
**Answer**: As granular as possible but don't be afraid to let the panel know you don't know.

**Question**: Is there a sandbox or dev environment available to learn GCP hands-on?
**Answer**: Click the try it out link and create your own environment.

**Question**: With reference to the demo, is this high-level or working? If working, where do we get our hands on sample data sets?
**Answer**: The expectation is that you create a working demo (it doesn’t have to be perfect). Create sample data yourself; there are a number of online tools you can use to generate mock data - just Google it.

## Case-Specific FAQs

### Scenario 4. Big Data: Aggregating and Analyze Doubleclick Data

**Question**: In the requirements description, it says “Doubleclick data is in a star schema.” The schema you pointed me to looks like a flat file. Is there a star schema that is already created that I can look into, or should I infer a star schema out of it?
**Answer**: Think very small star schema. There is a transfer table and a metadata table.

**Question**: In reading Preparing Data for Loading, it appears Google is strongly suggesting to denormalize table schemas. Would it be realistic to implement a logical star schema from #1 as a single physical table? Traditional dimensional model points to a physical star, but the Google docs seem to point to a single table.
**Answer:** Big Query works best with a single table as joins can start to slow it down. Consider providing star schema in BQ but materialize to a new table for performance.

**Question**: Since there is no Doubleclick dataset, would it be ok if I demonstrate the design principles on another time-series dataset?
**Answer:** Yes. Bonus if you want to create some data that looks similar to Doubleclick. There are a number of online tools you can use to generate mock data - just Google it.

**Question**: The requirements also state “not all data is 100% accurate and may be changed or added in subsequent daily dumps.” Late arrivals that need to change already-inserted data will need to be matched. Typically there is some kind of row identifier, whether it’s a row id or a combination of immutable columns (usually a combination of dimension columns). Is this a reasonable expectation?
**Answer:** It is not that a row of data changes, but new data shows up. I.e., an impression from Monday wasn't included until Tuesday's dump.

**Question**: The “some considerations” section asks about how to “extend the system to include data from other ad exchanges.” Is it reasonable to expect that dump files from the other exchange also contain similar pertinent information (whether it is in or not in the same format)?
**Answer:** It should be a similar format, but also think about how you might want to combine other systems (POS, ERP, etc.) to get more value from the data.

**Question**: What percentage of a given day’s data is late from previous days (late arrival)?
**Answer:** Assume 5-10% and just put that in your assumptions. Ideally it shouldn't really change the solution if it was 2% or 40%.

**Question**: How long does the ad agency need to retain the data?
**Answer:** Make a recommendation - how long should they retain it? What value do they get out of longer retention times?

**Question**: The requirements state “they also need to provide reporting and/or exports to individual clients on only their data.” Is the expectation for the ad agency’s clients to be able to access a reporting service and create their own report (as opposed to canned reports)?
**Answer:** First would be ability to see their own data, 2nd is ability to download their own data, 3rd is perhaps some more sophisticated way to analyze their data besides just canned and downloadable data.

**Question**: The requirements also asks “what type of reporting mechanisms are supported?” Is this referring to platforms e.g. Qlik, Tableau, Excel, etc.?
**Answer:** Yes. How could you build in reporting systems into the solution?

**Question**: Would it be reasonable to expect that the ad agency would want to simplifying the architecture by using as few disparate technologies?
**Answer:** That is an excellent goal, but use your judgement on when/why you should introduce new tech. I.e., don't try to make everything a nail if you have a hammer...

**Question**: Do you know if there is any example data to show the expected schema of the data being ingested? That would be needed for an end to end working prototype.
**Answer**: There is no example data. Most candidates create dummy data to demonstrate the concepts.