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: File Sharing & Management System
    Scenario 3: Real-Time Data Ingest: Traffic Telemetry
    Scenario 4: Networking - Cloud SSL Inspection
    Scenario 5: Web Serving: Autocomplete
    Scenario 6: 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. File Sharing & Management System

A large manufacturing firm which houses large number of teams - HR, legal, market research, product design, etc. - is looking to build a centralised file sharing and management system.

Being a manufacturing firm, they deal with large volumes of documents on a day-to-day basis: product designs, vendor and supplier contracts, go-to-market templates, compliance reports and so on. At present, most of the document management in the organization is manual. Basic tasks such as collaboration between employees within/outside groups, retrieving the most updated version of a document, maintaining basic audit trail on files are done either over emails or unmanaged groups. This not only puts a big dent on their operational efficiencies but also poses a big security threat.

The proposed system should have a simplistic web interface that allows users to login and upload/download/view files. Each user should have predefined role and access (defined by the group admin/super-user). For example, by default, somebody belonging to market research group should not have access to files/folders in legal services. The files can be shared across groups only if they do not violate any overarching policies set by group admins.

The logged in user should be able to see the list of files uploaded in the group to which he/she belongs. They can view versions, download and upload files. The user should also be able to search for files through a simple search bar.

Minimum Expectations

- Running prototype of the file sharing system on GCP and a walk through of components - choice of front end (GAE/GCE/GKE), choice of data storage (blobs and metadata); explain pros/cons of your choices.
- At least one alternative architecture with pros/cons
- Users belonging to different groups being isolated from each other by default with no sharing of documents allowed. Walk through of design implementation of how this isolation was achieved.
- Basic reporting for admins: Current and past user sessions and activities undertaken (files viewed/downloaded/uploaded etc.). Walk through of GCP components, their pros and cons
- Estimated running costs on GCP

Nice-to-Have

- Demonstrate complex file sharing and management rules. Setting group level policies that should be adhered to for sharing files/folders. For example, one of the policies can be - no file carrying employee equity grants under HR can be shared outside the group; other files are okay to be shared.
- Advanced analytics: track devices connected, third party access, white/black list ips, domains, specific users, track failed login/access attempts. This may not be built as part of the demo but a walkthrough of system design would be great - what GCP component(s) can be used and why?

Assumptions and Considerations

- To gauge the scale of this application, assume there are 10 teams with 1,000 users each. Each user uploads 2-3 documents, downloads ~10 documents every day. The frequency of search can vary randomly per user per day. The busiest times for this application is between 11am - 3pm every day. There is almost no activity expected from 7pm - 7am every day on typical days.
- You do not need to integrate with an existing identity management system (while it's nice to have, it's not a requirement). Start with a plain vanilla system.
- The considerations for the alternate design can be quicker turnaround on product development, ease of deployment/management, better costs, minimal learning curve for engineering etc.
- What can be done to optimise costs?

Scenario 3. 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 4. Networking - Cloud SSL Inspection

A major retail customer is testing Chrome OS as a replacement for their in-store point-of-sale and kiosk terminals in all of their stores (1000+ across North America). However, they are encountering issues due to firewall restrictions on their in-store network, which is PCI-compliant.

They would like to open up traffic on this network only to the hostnames given in this Chrome support article, without opening up traffic to all of Google. The Chrome devices are being used as point-of-sale (POS) machines, and they need to be locked down from access to G Suite services (e.g. Gmail, Drive) to maintain PCI compliance.

However, all traffic to and from Google happens over SSL, which breaks their web filtering since their aging network infrastructure does not support SSL inspection. All encrypted traffic to any Google endpoint appears in logs as destined for google.com, with no further detail. Again, for compliance reasons, general traffic to google.com cannot be allowed on this network.

It is necessary for the Chrome devices to have constant contact with Google's policy servers, so that policy changes made in the Chrome Management Console can reach the POS devices. The customer cannot ever connect the POS devices to a non-PCI network. Two-way traffic between the POS devices and the specific Google endpoints mentioned in the above article must be allowed.

They do maintain a specific appliance that communicates with their external payment gateway over SSL, but this device is not designed to handle general routing - it's an appliance they bought from their payment gateway vendor. Setting up an internal SSL proxy in their own environment is not a desirable option for them - the retailer is actively moving away from on-premise infrastructure for cost reasons, and an SSL proxy appliance seems like too much of a hassle to deploy for such a limited use-case.

The customer vaguely knows about Google Cloud Platform, and wants to know if it's possible to deploy an SSL proxy to GCP, connect it securely to their on-premise network, and use it to filter traffic to the desired Google endpoints. PCI-DSS requires that every component in the Cardholder Data Environment (CDE - every part of the network that is actually handling payment data, starting with the Chrome POS machine) remain isolated and segregated from the internet and even other internal networks. This means that traffic moving from on-premise to the proxy cannot be exposed to an external network at any point in between unless that connection is segregated by a firewall.

Minimum Expectations

- One recommended architecture with details on how the flows through the system and descriptions of each component
- Prototype(s) running on GCP
- At least one alternative architecture with pros/cons

Some Considerations

- Please see the FAQs for PCI-DSS 1.*.* requirements to specifically consider for this case
- The customer has provided you with this diagram of their current network
- There are 1000 stores with 3-5 POS machines in each. The system must be able to scale and handle that sort of load.
- It is not necessary to demo this solution with a Chrome device as the client. However, be sure you can prove that all of the above requirements are effectively met regardless of the client device that you use.
- Does the customer need to use a VPN? Is there any other way to extend the on-premise environment to the cloud? Which do you recommend and why?
- It likely won't be practical to 100 percent emulate the customer's environment in the demo, but be able to explain anything that will differ in a production environment for compliance sake.

Scenario 5. 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 6. 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 2: File Sharing and Management System

Question: Do I need to integrate with an existing identity management system?
Answer: This is nice to have, you don't necessarily have to. Start with a plain vanilla system.

Scenario 4: Networking - Cloud SSL Inspection

PCI-DSS 1.*.* Requirements to specifically consider:

This is not the full list of requirements for compliance with PCI-DSS Requirement 1, but the ones that you specifically should think about as you propose changes to this system. The below will be the responsibility of Google to show how compliance is maintained in a hybrid-cloud environment.

- Formal process for approving all network connections and changes to the firewall (1.1.1)
    - *Think about:* How can you limit access to network controls on the Google Cloud side to appropriate users? How can this fit into the customer's existing change control processes?
- Current network diagram with all connections to cardholder data, including wireless (1.1.2)
    - Customer has provided this.
- Requirements for a firewall at each internet connection and between the demilitarized zone (DMZ) and the internal network zone (1.1.4)
    - *Think about:* Are there points where your system will interact with another network, or with the internet?
- Description of groups, roles, and responsibilities for logical management of network components (1.1.5)
    - *Think about:* Assume that the customer already has this, and just wants to integrate their existing management structure with Google Cloud Platform.
- Documentation and business justification for use of all services, and ports allowed, including documentation of security features implemented for those protocols considered to be insecure
    - *Think about:* Be able to justify every component you are adding to the system, why it's important, and how it's secured and compliant.
- Restrict inbound and outbound traffic to that which is necessary for the cardholder data environment, and specifically deny all other traffic. (1.2.1)
- Implement a DMZ to limit inbound traffic to only system components that provide authorized publicly accessible services, protocols, and ports (1.3.1)
- Limit inbound Internet traffic to IP addresses within the DMZ (1.3.2)
- Do not allow any direct connections inbound or outbound for traffic between the Internet and the cardholder data environment (1.3.3)
- Implement anti-spoofing measures to detect and block forged source IP addresses from entering the network (1.3.4)
    - *Think about:* How do you ensure that inbound traffic is actually coming from Google and not from an attacker masquerading as Google?
- Do not allow unauthorized outbound traffic from the cardholder data environment to the Internet (1.3.5)
- Implement stateful inspection, also known as dynamic packet filtering (1.3.6)
    - *Think about:* How can you limit inbound connections from specified Google endpoints to only those that are replies to requests from client machines?

Scenario 6. 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.