☑ 13. GCP — 50 Preguntas Técnicas (Respuestas Completas)

---

1. Servicios principales de datos en GCP

- BigQuery
- Cloud Storage
- Pub/Sub
- Dataflow
- Dataproc
- Cloud Composer (Airflow)
- Data Fusion
- Bigtable
- Firestore
- Looker / Looker Studio

---

2. BigQuery vs Cloud SQL vs Firestore

Servicio	Tipo	Uso
BigQuery	DWH serverless	Analítica masiva
Cloud SQL	Base relacional administrada	OLTP pequeño/mediano
Firestore	NoSQL (documentos)	Apps web/móvil, datos semi-estruc.

---

3. ¿Qué es BigQuery?

Data Warehouse *serverless* para analítica a gran escala con SQL.

---

4. Almacenamiento columnar en BigQuery

Guarda los datos por columna (no por fila) → lecturas rápidas, compresión, escaneo reducido.

---

5. Dataset en BigQuery

Contenedor lógico de tablas, vistas y rutinas.

---

6. Tabla particionada

Tabla dividida en segmentos basados en fecha, timestamp o enteros → mejora costo y rendimiento.

---

7. Particionamiento por tiempo vs rango

- **Tiempo:** DATE, TIMESTAMP datetime.
- **Rango:** valores numéricos (ids, precios).

---

8. Tabla clusterizada

Organiza datos por columnas clave internamente para leer solo bloques relevantes.

---

9. Cloud Storage

Almacenamiento de objetos (archivos) escalable y durable.

---

10. Estándar vs Nearline vs Coldline vs Archive

Tipo	Costo almacenamiento	Costo acceso	Uso
Standard	medio	bajo	acceso frecuente
Nearline	bajo	medio	acceso mensual
Coldline	menor	alto	acceso trimestral
Archive	mínimo	muy alto	acceso anual

---

11. Pub/Sub

Sistema de mensajería *event-driven*. Conecta productores y consumidores en pipelines streaming.

---

12. Dataflow

Servicio serverless para procesamiento batch o streaming basado en Apache Beam.

---

13. Dataflow batch vs streaming

- **Batch:** datasets completos, mayor latencia.
- **Streaming:** eventos en tiempo real, baja latencia.

---

14. Dataproc

Cluster administrado de Hadoop/Spark. Para migrar workloads existentes o ETLs Spark.

---

15. Dataflow vs Dataproc

- **Dataflow:** serverless, autoscale, pipelines Beam.
- **Dataproc:** Spark/Hadoop tradicional; infraestructura más controlable.

---

16. Cloud Composer

Orquestador basado en Apache Airflow completamente administrado.

---

17. Composer vs Dataflow

- **Composer:** *orquesta*, schedule, dependencias.
- **Dataflow:** *procesa* datos.

---

18. BigQuery ML

Permite entrenar y ejecutar modelos ML con SQL directamente en BigQuery.

---

19. ML directamente en BigQuery

`CREATE MODEL`, `ML.PREDICT`, `ML.EVALUATE`, `ML.EXPLAIN`.

---

20. Looker Studio

Herramienta BI serverless para dashboards; se conecta directamente a BigQuery.

---

21. IAM

Sistema de control de acceso basado en identidades, roles y políticas.

---

22. Roles predefinidos vs personalizados

- **Predefinidos:** específicos del servicio, finos.
- **Personalizados:** configurados por el usuario.

---

23. VPC

Red privada virtual para controlar tráfico interno, subredes, firewall.

---

24. Service Account

Identidad para servicios; usada para acceso a APIs y permisos específicos.

---

25. Cloud Functions

Función serverless para eventos pequeños (ETL ligero, triggers).

---

26. Cloud Run vs Cloud Functions

Cloud Run	Cloud Functions
Ejecuta contenedores	Ejecuta funciones
Más flexible	Más simple
HTTP-based	Event-based

---

27. Serverless vs VM

- **Serverless:** no se administra infraestructura.
- **VM (Compute Engine):** control total, pero más mantenimiento.

---

28. Cloud Logging

Servicio centralizado para logs de aplicaciones, GCP, audit logs.

---

29. Cloud Monitoring

Recolecta métricas, alertas y dashboards para servicios GCP.

---

30. Asegurar datos sensibles en BigQuery

- Column-level security
- Row-level security
- Masking
- IAM granular
- DLP

---

31. CMEK

Claves de cifrado administradas por el cliente.

---

32. CMEK vs CSEK

- **CMEK:** clave en KMS; administrada por cliente.
- **CSEK:** clave externa proporcionada por el cliente durante la petición.

---

33. Manejar costos en GCP

- Limitar escaneos en BigQuery
- Particionar + clusterizar
- Monitorizar
- Alerts
- Precios flat-rate
- Lifecycle rules en Storage

---

34. Reservations y slots en BigQuery

- **Slots:** unidades de cómputo.
- **Reservations:** asignación estática de slots a proyectos o equipos.

---

35. Flat-rate vs on-demand

- **Flat-rate:** costo fijo mensual.
- **On-demand:** pagas por TB escaneado.

---

36. Exportar BigQuery → Cloud Storage

`EXPORT DATA` o desde UI.

---

37. Importar Cloud Storage → BigQuery

`LOAD DATA`, autocreate schema o especificado.

---

38. BigQuery BI Engine

Capa en memoria para acelerar dashboards (Looker/Looker Studio).

---

39. BI Engine vs columnar storage

- BI Engine acelera *cache en memoria*.
- Columnar es almacenamiento nativo de BigQuery.

---

40. Versionar pipelines Dataflow/Composer

- Git
- Images de contenedor versionadas
- Variables de entorno
- DAGS versionados

---

41. Cloud Data Catalog

Catálogo unificado de metadatos → linaje, descripciones, etiquetas.

---

42. Auditar accesos

- Audit Logs
- BigQuery Access Logs
- IAM Recommender
- Cloud Logging

---

43. Cloud DLP

Detecta y anonimiza datos sensibles: PII, tarjetas, patrones.

---

44. Labels en GCP

Metadatos para recursos (costo, dueño, equipo); útiles para gobernanza y billing.

---

45. GCP vs AWS en datos

AWS	GCP
Redshift	BigQuery
S3	Cloud Storage
Glue	Dataflow/Dataproc
Kinesis	Pub/Sub
EMR	Dataproc

---

46. Monitorear pipelines ETL en GCP

- Logs (Composer, Dataflow)
- Metrics en Monitoring
- Alertas
- Pub/Sub dead-letter

---

47. Custom metrics

Métricas personalizadas enviadas a Monitoring desde apps o pipelines.

---

48. Soluciones escalables

- Serverless primero
- Autoscaling
- Diseño idempotente
- Desacoplar con Pub/Sub
- Clusterizar BigQuery
- Multi-zone

---

49. Pub/Sub Lite

Versión económica de Pub/Sub con costos predecibles; menor SLA.

---

50. Buenas prácticas arquitectura de datos GCP

- Usar BigQuery particionado/clusterizado
- Desacoplar con Pub/Sub
- IaC con Terraform
- Seguridad IAM mínima
- Monitoreo y alertas
- Lineage + catalog
- Data quality gates
- Separación de ambientes