Gcp Professional Data Engineer Cheat Sheet

3.1. Cloud Storage (GCS)

- Blob storage. Upload any bytes to a location. The content **isn't indexed** at all just stored. (Amazon S3)
- Virtually **unlimited** storage
- Nearline and Coldline: for ~1 sec lookup for access, charged for volume of data accessed
  - **Nearline** for once per month
  - **Coldline** for once per year
- buckets to segregate storage items
- geographical separation:
  - persistant, durable, replicated
  - spread data across zones to minimise impact of service disruptions
  - spread data across regions to provide global access to data
- Ideal for storing but **not for high volume of read/write** (e.g. sensor data)
- A way to store the data that can be **commonly used by Dataproc and BigQuery**

Encryption:
- Google Cloud Platform encrypts customer data stored **at rest** by default
- Encryption Options:
  - Server-side encryption:
    - **Customer-supplied encryption keys**: You can create and manage your own encryption keys for server-side encryption
    - **Customer-managed encryption keys**: You can generate and manage your encryption keys using **Cloud Key Management Service**.
  - Client-side encryption: encryption that occurs before data is sent to Cloud Storage.

3.2. Cloud SQL & Spanner

- Managed/No ops relational database (MySQL and PostgreSQL) like **Amazon RDS**.
- best for **gigabytes** of data with **transactional** nature
  - Low latency
  - Doesn't scale well beyond GB's
  - data structures and underlying infrastructure is required.
- **Spanner is a distributed and scalable solution** for RDBMS however also **more expensive**.
- Management:
  - Managed backups & automatic replication
  - fast connection with GCE/GAE
  - uses Google security
  - Flexible pricing, pay for when you use it

3.3. BigTable

Feature:
- Stored on Google's internal store **Colossus**
- **no transactional** support (so can handle **petabytes** of data)
- **not relational** (No SQL or joins), ACID only at row level.
  - Avoid schema designs that require atomicity across rows
- **high throughput**: Throughput has linear growth with node count if correctly balanced.
- work with it using **Hbase API**
- **no-ops**, auto-balanced, replicated, compacted.

Query:
- **Single key lookup**. No property search.
- Stored **lexicographically** in big endian format so keys can be anything.
- **quick range lookup**

Performance:
- Fast to **petabyte** scale, **not a good** solution for storing **less than 1 TB of data**.
- **Low-latency** read/write access
- High-throughput analytics
- **Native time series** support
- for large **analytical and operational** workloads
- designed for **sparse tables**

Key Design:
- Design your keys **how you intend to query**.
- If your most common query is the **most recent data**, use a **reverse date stamp** at the end of the key.
- Ensure your **keys are evenly distributed** to void **hot spotting**. This is why date stamps as a key or starting a key is bad practise as all of the most recent data is being written at the same time.
  - For **historical data analytic, hotspot issue** may not a biggest concern.
- For **time-series** data, use **tall/narrow** tables. Denormalize- prefer multiple tall and narrow tables
- **avoid hotspotting**
  - **Field promotion (preferred)**: Move fields from the column data into the row key to make writes non- contiguous.
  - **Salting**: (only where field promotion does not resolve) Add an additional calculated element to the row key to artificially make writes non- contiguous.

Performance Test
- learns about your **access patterns** and will adjust the metadata stored in nodes in order to try balance your workloads.
- This takes minutes to hours and requires to use at least **300GB of data**
- use a **production** instance
- Stay **below the recommended storage utilization** per node.

GCP Professional Data Engineer Cheat Sheet: Preparing for the Google Cloud Platform Professional Data Engineer certification can be a daunting task, but having a cheat sheet can significantly streamline your study process. This guide will provide an overview of essential topics, concepts, tools, and best practices that you need to understand in order to excel in the exam and in your data engineering career on GCP.

Understanding the Role of a Data Engineer

Data engineers are responsible for designing, building, and maintaining the infrastructure and systems that allow organizations to collect, store, and analyze data. Their primary focus is to ensure that data flows seamlessly across various systems and is readily available for analysis by data scientists and business analysts.

Key Responsibilities

1. Data Modeling: Designing efficient data models that support business use cases.
2. Data Pipeline Development: Building ETL (Extract, Transform, Load) processes to move data between systems.
3. Data Storage Solutions: Selecting appropriate storage options (e.g., BigQuery, Cloud Storage) based on user needs.
4. Performance Monitoring: Implementing monitoring systems to ensure data

pipelines run efficiently.
5. Collaboration: Working with data analysts and data scientists to understand data requirements.

Core Concepts in GCP Data Engineering

Understanding key concepts in Google Cloud Platform is crucial for any professional data engineer. Below are the core areas you should focus on:

1. Google Cloud Storage (GCS)

- Used for storing unstructured data.
- Supports various data formats such as JSON, Parquet, Avro, etc.
- Provides features like versioning, lifecycle management, and access control.

2. BigQuery

- A fully managed data warehouse that allows for SQL queries on large datasets.
- Supports real-time analytics and can handle structured and semi-structured data.
- Features such as partitioning, clustering, and materialized views enhance performance.

3. Dataflow

- A serverless data processing service that allows for batch and stream processing.
- Supports Apache Beam for building data processing pipelines.
- Automatically scales to handle varying workloads.

4. Pub/Sub

- A messaging service for building event-driven architectures.
- Supports message queuing and real-time data streaming.
- Allows for decoupling of services, improving system resilience.

5. Dataproc

- A managed Spark and Hadoop service for big data processing.
- Facilitates the running of batch processing jobs using familiar open-source tools.
- Supports integration with GCS and BigQuery.

Data Engineering Workflow on GCP

Understanding the workflow of a data engineering project is vital. Here’s a typical data pipeline flow in GCP:

1. Data Ingestion:
- Use Pub/Sub for streaming data.
- Use Data Transfer Service for batch data.

2. Data Storage:
- Store raw data in Google Cloud Storage.
- Use BigQuery for structured data storage and analytics.

3. Data Processing:
- Use Dataflow for real-time or batch processing.
- Use Dataproc for processing data with Hadoop or Spark.

4. Data Analysis:
- Use BigQuery for running SQL-based analytics.
- Use Looker or Data Studio for visualization.

5. Data Monitoring and Management:
- Utilize Cloud Monitoring and Logging to track performance.
- Automate tasks with Cloud Functions or Cloud Run.

Best Practices for GCP Data Engineering

Following best practices ensures that your data engineering projects are efficient, scalable, and secure. Here are some important best practices:

1. Optimize Data Storage

- Choose the right storage options based on the nature of the data (e.g., structured vs. unstructured).
- Implement data partitioning and clustering in BigQuery to improve query performance.
- Use lifecycle management to delete or archive data that is no longer

needed.

2. Design Scalable Pipelines

- Use serverless services like Dataflow for auto-scaling based on workload.
- Implement modular pipelines to isolate different processing steps.
- Use Pub/Sub for decoupling components and enabling event-driven architecture.

3. Ensure Data Quality

- Implement data validation checks during data ingestion.
- Create alerts for monitoring data anomalies using Cloud Monitoring.
- Regularly audit and clean datasets to remove duplicates and inaccuracies.

4. Secure Your Data

- Use Identity and Access Management (IAM) to control access to resources.
- Enable encryption for data at rest and in transit.
- Regularly audit permissions and access logs.

Tools and Technologies in GCP

Familiarity with various tools and technologies is essential for a GCP data engineer. The following are some of the most important tools you should know:

1. Google Cloud Console

- A web-based interface for managing GCP resources.
- Provides dashboards for monitoring and managing services.

2. Cloud SDK and gcloud Command-Line Tool

- Command-line interface for managing GCP resources programmatically.
- Useful for automation and scripting.

3. Cloud Functions

- A serverless execution environment for running small bits of code in response to events.
- Ideal for lightweight data processing tasks.

4. Cloud Run

- A managed compute platform for running containers.
- Combines the benefits of containerization with serverless architecture.

Exam Preparation Tips

Preparing for the GCP Professional Data Engineer exam requires a strategic approach. Here are some tips to help you succeed:

1. Review the Exam Guide: Familiarize yourself with the topics covered in the exam by reviewing the official exam guide provided by Google.

2. Hands-on Practice: Create a GCP account and work on real-world projects to gain practical experience. Utilize free tier services for cost-effective learning.

3. Online Courses and Practice Tests: Use resources from platforms like Coursera, Pluralsight, or A Cloud Guru to take structured courses. Practice tests can help you gauge your readiness.

4. Join Study Groups: Engage in online communities or forums such as Google Cloud Community or Reddit to discuss topics and share knowledge with peers.

5. Stay Updated: The field of data engineering is ever-evolving. Keep up with the latest updates and features in GCP by following Google Cloud blogs and attending webinars.

Conclusion

The GCP Professional Data Engineer Cheat Sheet serves as a valuable resource for anyone looking to enhance their understanding of data engineering on Google Cloud Platform. By mastering the core concepts, tools, best practices, and workflows, you will be well-prepared not only for the certification exam but also for a successful career in data engineering. As you embark on this journey, remember that hands-on experience and continuous learning are key to your success in this dynamic field.

Frequently Asked Questions

What is the purpose of the GCP Professional Data Engineer certification?

The GCP Professional Data Engineer certification validates an individual's ability to design, build, operationalize, secure, and monitor data processing systems in Google Cloud Platform.

What are the key areas covered in the GCP Professional Data Engineer exam?

The exam covers areas such as data ingestion, data storage, data processing, data analysis, machine learning, and operationalizing data processing systems.

What is a cheat sheet and how can it help in preparing for the GCP Professional Data Engineer exam?

A cheat sheet is a concise set of notes that summarize key concepts and terms, helping candidates quickly review important information and study efficiently for the exam.

What tools and services are essential to know for the GCP Professional Data Engineer certification?

Key tools and services include BigQuery, Cloud Dataflow, Cloud Dataproc, Cloud Pub/Sub, Cloud Storage, and Google Cloud AI and Machine Learning services.

How often is the GCP Professional Data Engineer exam updated?

The GCP Professional Data Engineer exam is periodically updated to reflect changes in technology and best practices, typically every 1-2 years.

What is the format of the GCP Professional Data Engineer exam?

The exam consists of multiple-choice and multiple-select questions, and it is administered online with a time limit of 2 hours.

Are there any recommended resources for studying for the GCP Professional Data Engineer exam?

Recommended resources include the official Google Cloud training courses,

Qwiklabs, practice exams, and documentation on Google Cloud services.

What is the passing score for the GCP Professional Data Engineer exam?

The passing score for the GCP Professional Data Engineer exam is not publicly disclosed, but it is generally understood to be around 70%.

Gcp Professional Data Engineer Cheat Sheet

Google Cloud Platform
Google Cloud Platform lets you build, deploy, and scale applications, websites, and services on the same ...

Google Cloud console
Google Cloud Console has failed to load JavaScript sources from www.gstatic.com.

Google Cloud Platform
Google Cloud Platform enables you to build, deploy, and scale applications using Google's infrastructure.

Cloud 관리 콘솔 | *Google Cloud*
투명한 가격 책정 방식으로 비용 절감 Google Cloud는 사용한 만큼만 지불하는 가격 책정 방식으로 월별 사용량과 선불 리소스의 할인율을 ...

云管理控制台 | **Google Cloud**
通过我们直观易用的控制台界面，快速上手使用。此外，您还可以免费使用超过 20 种产品（不超过每月限额）。

Google Cloud Platform
Google Cloud Platform lets you build, deploy, and scale applications, websites, and services on the same infrastructure as Google.

Google Cloud console
Google Cloud Console has failed to load JavaScript sources from www.gstatic.com.

Google Cloud Platform
Google Cloud Platform enables you to build, deploy, and scale applications using Google's infrastructure.

Cloud 관리 콘솔 | **Google Cloud**
투명한 가격 책정 방식으로 비용 절감 Google Cloud는 사용한 만큼만 지불하는 가격 책정 방식으로 월별 사용량과 선불 리소스의 할인율을 기준으로 자동 할인을 제공합니다. 지금 Google에 문의하여 견적을 받아보세요.

云管理控制台 | Google Cloud
通过我们直观易用的控制台界面，快速上手使用。此外，您还可以免费使用超过 20 种产品（不超过每月限额）。

Console de gestion cloud | Google Cloud
Réalisez une démonstration de faisabilité sans frais avec 300 $ de crédits offerts et essayez plus de 150 produits dans la console Google Cloud.

Google Cloud console
O console do Google Cloud não conseguiu carregar fontes JavaScript a partir de www.gstatic.com. Os possíveis motivos são: www.gstatic.com ou os respectivos endereços IP foram bloqueados pelo administrador da sua rede O Google bloqueou temporariamente sua conta ou rede devido a um excesso de solicitações automatizadas Fale com o administrador ...

Google Cloud console
利用 Google Cloud Marketplace 智能掌控支出，享受更快速的采购体验，彻底告别 Google Cloud 承诺支出。浏览目录，其中有 2000 种以上的 SaaS、虚拟机、开发技术栈和 Kubernetes 可供选择，它们都已经过优化，在 Google Cloud 上运行时可达到最佳效果。

クラウド管理コンソール **| Google Cloud**
$300 分の無料クレジットを使用して概念実証を無料で構築し、Google Cloud コンソールで 150 以上のプロダクトを試すことができます。

Google Cloud console
Spend smart, procure faster and retire committed Google Cloud spend with Google Cloud Marketplace. Browse the catalog of over 2000 SaaS, VMs, development stacks, and Kubernetes apps optimized to run on Google Cloud.