ChatGPT

Data Ingestion Engineer: Step 1 - ReadMe Guide

This document provides a practical guide for Step 1 of a real-world cloud data ingestion pipeline, especially from on-prem SQL databases like Microsoft SQL Server or PostgreSQL, into platforms like BigQuery using tools such as Dataproc or Dataflow.

Step 1: Understanding the Data Ingestion Requirements

🎮 Goal

To gather all necessary information to design a robust, scalable, and secure ingestion pipeline.

Tasks in Step 1

1. Identify Data Sources

- **Where is the data coming from?**
  Examples: APIs, on-prem databases (PostgreSQL, SQL Server), files in Cloud Storage, Kafka, etc.

- **What formats are involved?**
  Examples: CSV, JSON, Parquet, Avro, ORC

- **Is the data structured, semi-structured, or unstructured?**

2. Understand the 3 Vs of Big Data

- **Volume**: How much data is expected (daily/hourly)?
- **Velocity**: Is it batch, near real-time, or streaming?
- **Variety**: Different schemas, formats, or sources involved?

3. Define Frequency and Latency Requirements

- How often should data be ingested? (e.g., hourly, daily, real-time)
- What is the acceptable delay from source to destination?

4. Determine Security and Access Needs

- Are there sensitive fields or compliance constraints (PII, HIPAA)?
- What IAM roles/service accounts are required?
- Is encryption needed (in transit or at rest)?

5. Profile the Data (Optional but Important)

- What do sample records look like?
- Are there missing fields, anomalies, or consistency issues?

Output Checklist for Step 1

- [x] Clear list of data sources and formats
- [x] Ingestion frequency and volume estimates
- [x] Destination targets (e.g., BigQuery datasets)
- [x] Security/IAM/Compliance considerations
- [x] Suggested architecture: Dataproc, Dataflow, or hybrid

Common Database Port Numbers

Ensure you can reach your source systems over the appropriate network ports:

Database System	Default Port
Microsoft SQL Server	`1433`
PostgreSQL	`5432`
MySQL	`3306`
Oracle DB	`1521`
MongoDB	`27017`
Kafka Broker	`9092`
Redis	`6379`
Google Cloud SQL (Postgres)	`5432`
BigQuery	N/A (API)

Note: When working with cloud VMs or Dataproc clusters, ensure firewall rules and VPC peering configurations permit traffic on these ports.

📊 Example Scenario

Requirement	Example
Source	On-prem PostgreSQL database (`5432`)

Requirement	Example
Frequency	Daily at midnight
Target	BigQuery sales_dataset
Method	Apache Spark job on Dataproc
Volume	~10 GB/day
Security	SSL encryption, VPC peering, IAM roles for service acct

---

🏆 Best Practices

- Use **Cloud VPN** or **Interconnect** for secure hybrid cloud connectivity
- Test DB connections using CLI tools (e.g., `psql`, `sqlcmd`)
- Use **Secret Manager** or encrypted environment variables for credentials
- Maintain ingestion logs for observability and auditing
- Automate schema validation and data quality checks

To Do Next

- Implement architecture using IaC (e.g., Terraform for Dataproc + Cloud Storage)
- Create airflow/Dataform pipelines for orchestration
- Monitor ingestion performance and errors with Cloud Monitoring

---

> This README serves as a base documentation for your GitHub repository covering cloud-based data ingestion pipelines. Adapt according to project specifics.