Google Cloud

Strategies and tools for efficient data ingestion

So far, you've learned that data ingestion is the process of collecting data from different sources and moving it to a staging area in the data pipeline. Data ingestion is an important concept in data analytics and processing. As an analyst, knowing how data is ingested, and the methods involved in this process, will give you a deeper understanding of data handling and management. In this reading, you'll learn more about the different data ingestion methods.

Data ingestion methods

Data ingestion can happen in one of two ways: batch or streaming. And, each of these methods dictate their own processing frequency.

With batch ingestion, data is collected over time and then processed into groups, also known as batches. Batch ingestion is an ideal method for ingesting data that doesn't need to be processed in real-time. For example, data for monthly billing reports, inventory processing, and subscription cycles can be processed monthly, quarterly, or annually.

With streaming ingestion, data is moved to destination storage as soon as it becomes available. In contrast to the batch method, streaming ingestion is optimized for data that must be processed in real-time. For example, banking apps and fraud detection systems rely on streaming to help provide their users with real-time alerts.

Configurations in data ingestion

Being aware of the challenges and precautions in data ingestion can help you prepare for issues that can happen on the job. When choosing between the batch or streaming method for a project, consider your data needs and business constraints to help you determine the ideal frequency of ingestion.

The quality and format of your data before ingestion is critical, and the condition of your data impacts the data journey at every stage. Data that is incorrectly formatted beginning at ingestion will need to be fixed in the transformation stage. Incorrect formats can result in lost or misrepresented data, leading to misleading or inaccurate analysis. Analysts can reduce quality and formatting issues by checking the data before ingestion, or implementing regular audits during the data journey.

If your data isn't configured correctly, errors will be present in the data pipeline, leading to discrepancies in your data. As an analyst, it's your responsibility to validate and verify the

integrity of your data. As a reminder, different data types and projects will have their own unique constraints that must be cross-checked during the data configuration process. Even though data will go through the same steps in the journey, these steps will look different for each project.

Pro tip: Always run a small test batch before fully implementing a new data ingestion process. Running a test batch can help you address potential issues earlier, and on a smaller scale.

Common tools for data ingestion

There are various tools to help facilitate the data ingestion process. Here's a list of common software that organizations can use to help with high-volume, real-time streaming data ingestion:

- Google Cloud Dataflow
- Apache Kafka, Apache NiFi, and Apache Beam
- Amazon Kinesis

Note: Different organizations have their own preferences for tools to use for data ingestion.

Key takeaways

Data ingestion can be done in either batch or streaming mode, depending on the need for real-time processing. Ensuring data quality, correct formatting, and configuration is important to prevent data loss or errors. As a data analyst, there are a variety of tools available for data ingestion that can help you facilitate the process, and avoid problems.

Resources for more information

Use the following resource to explore the world of data ingestion, its challenges, and best practices:

- Click this link for a brief overview of Big Data and the first layer of a data-gathering solution: Big Data Ingestion