Partitioning

Schedule partitioning

Not only are schedules great for reducing the amount of data our pipelines have to process, but they also help us guarantee that we can meet timing guarantees that our data consumers may need.

Logical partitioning

Conceptually related data can be partitioned into discrete segments and processed separately. This process of separating data based on its conceptual relationship is called logical partitioning. With logical partitioning, unrelated things belong in separate steps. Consider your dependencies and separate processing around those boundaries.

Also worth mentioning, the data *location* is another form of logical partitioning. For example, if our data is stored in a key-value store like Amazon's S3 in a format such as: `s3://<bucket>/<year>/<month>/<day>` we could say that our date is logically partitioned by time.

Size Partitioning

Size partitioning separates data for processing based on desired or required storage limits. This essentially sets the amount of data included in a data pipeline run. Size partitioning is critical to understand when working with large datasets, especially with Airflow.

NEXT