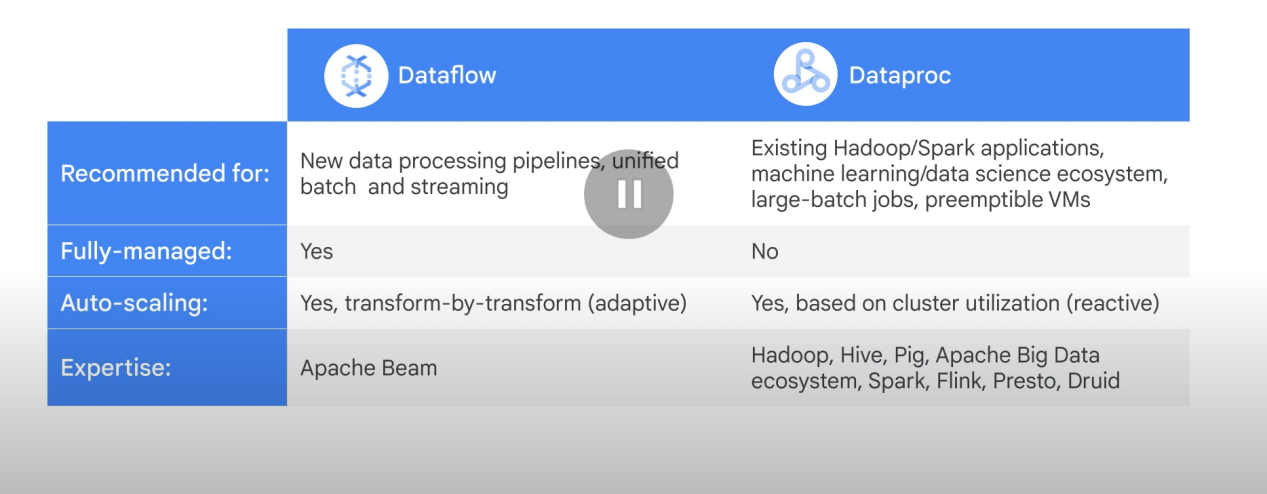
DataFlow provides an effective execution mecanism for **Apache Beam**

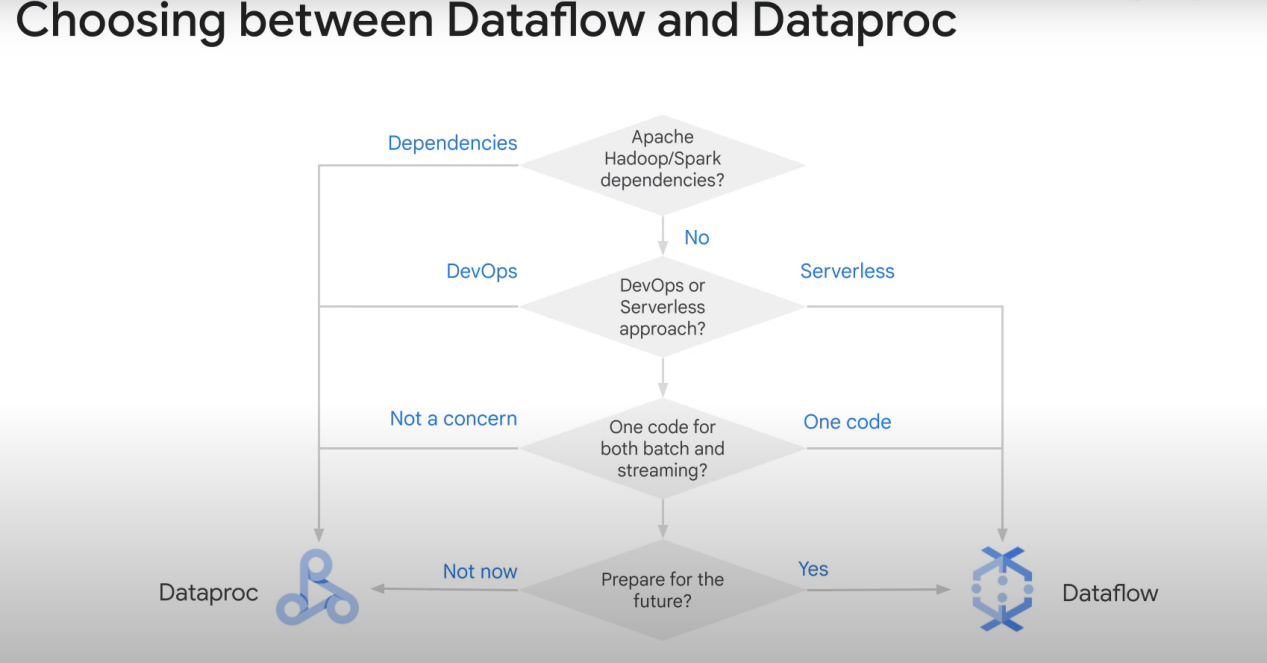
**You Can use Dataflow for ETL.**



**Dataflow is better than dataproc. You would only use dataproc if you have an exisiting project in Hadoop.**

\* Dataflow allows you to use the same code for both batch and streaming (real-time) data processing.

\* Dataflow is serverless: no need to manage clusters manually.

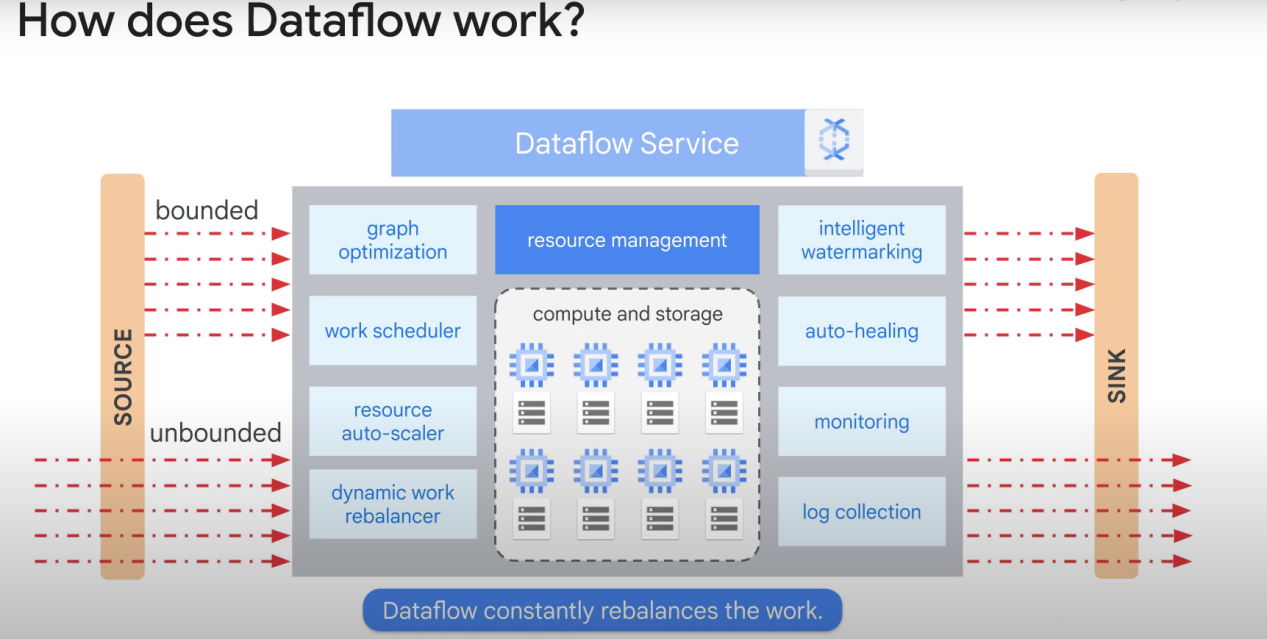


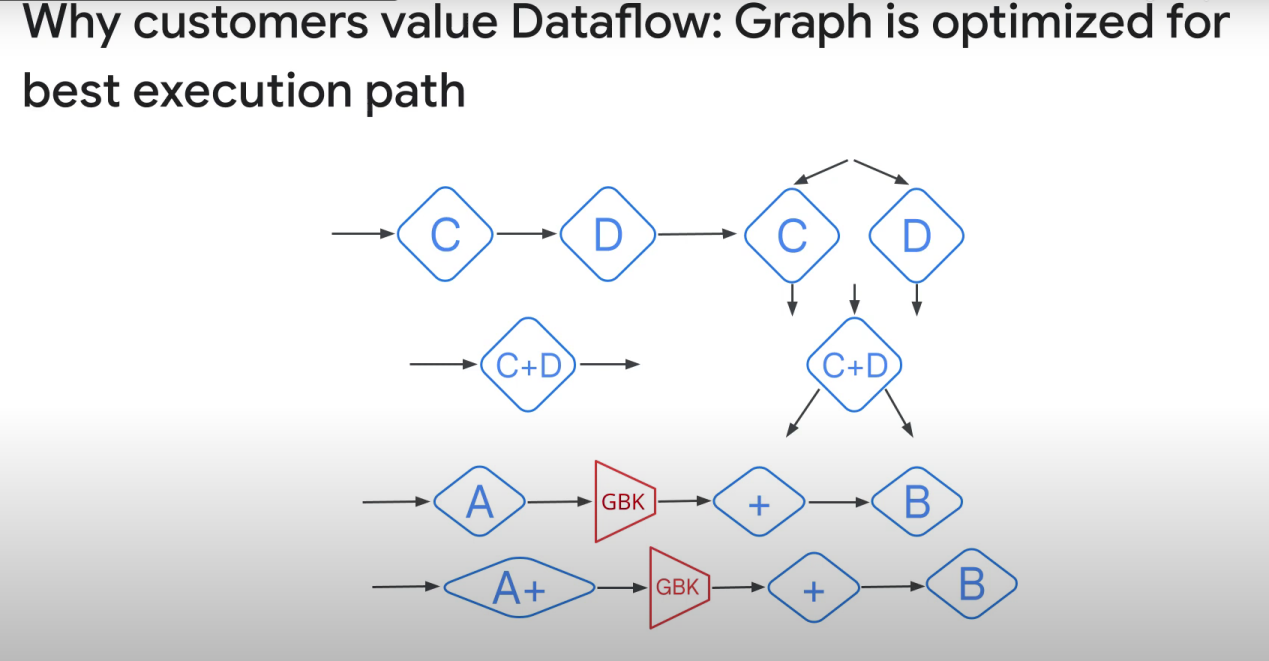
*The Beam Pipeline (of apache beam) precises what needs to be done*

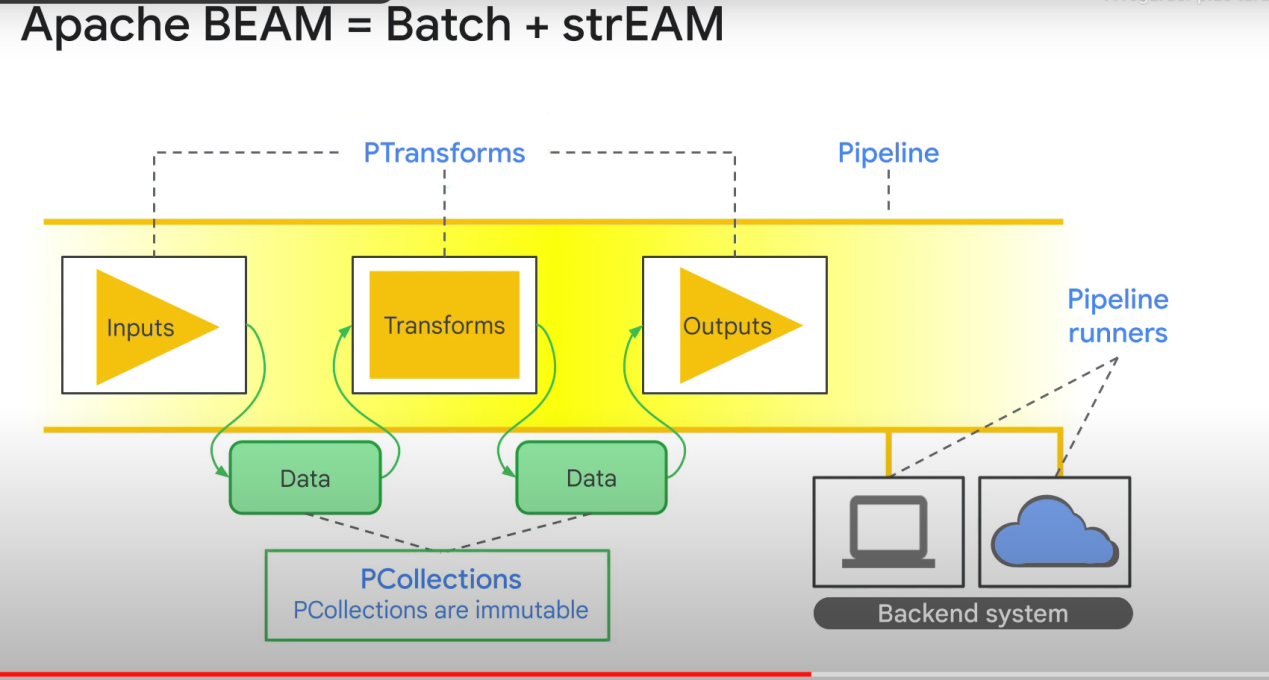
*VS*

*Dataflow chooses how to run the pipeline.*

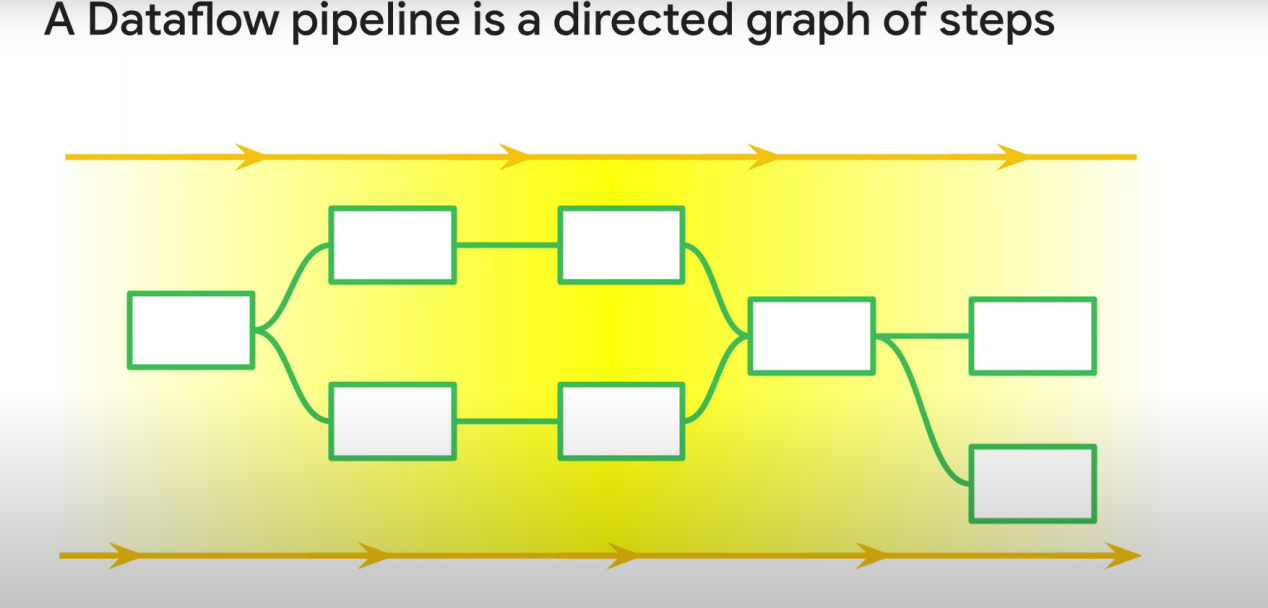
The pipeline contain steps like reading data, transforming it, and writing it.

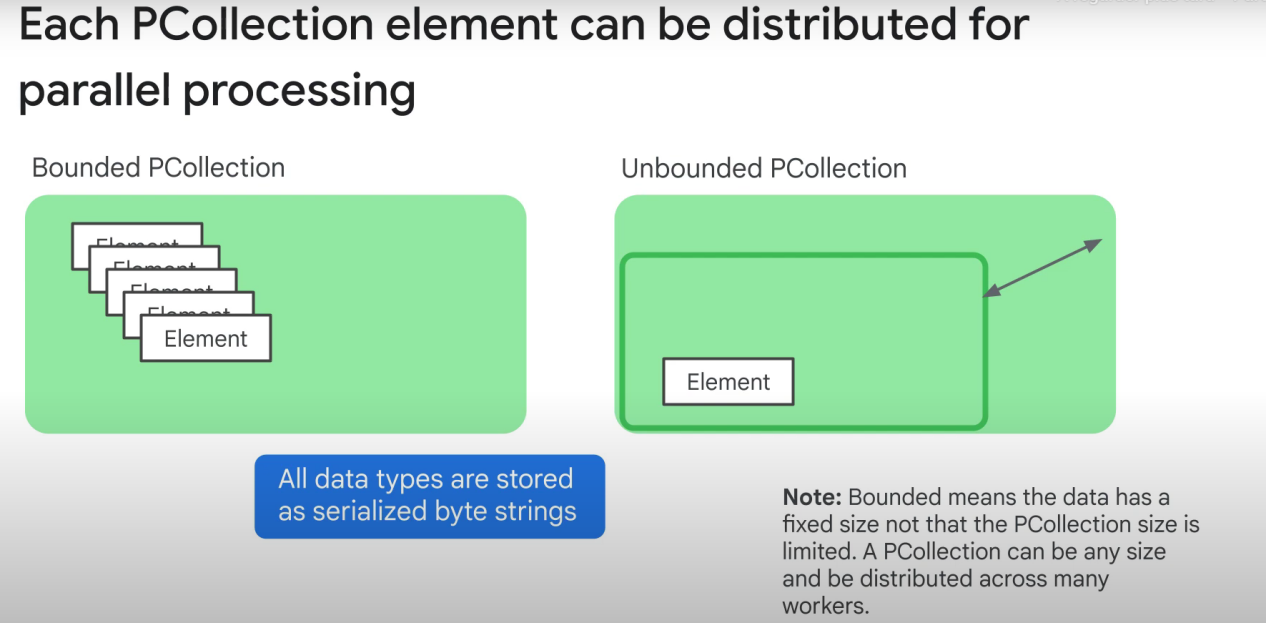
**Dataflow will optimize everything** (E.g: merge transformations together, breaking a job into units of work and schedule each unit to workers, auto-scale everything, dynamic work rebalance, etc...) 





DATA IS IMMUTABLE! WE CREATE A NEW COLLETION AFTER EACH PTRANSFORM. WE DO NOT CHANGE THE ORIGINAL COLLECTION.





All elements from PCollections are stored as **serialized byte strings**. We only need to deserialize them when doing PTransforms. **When sending the data via network, this saves us time** as we don’t need to serialize them before sending them.