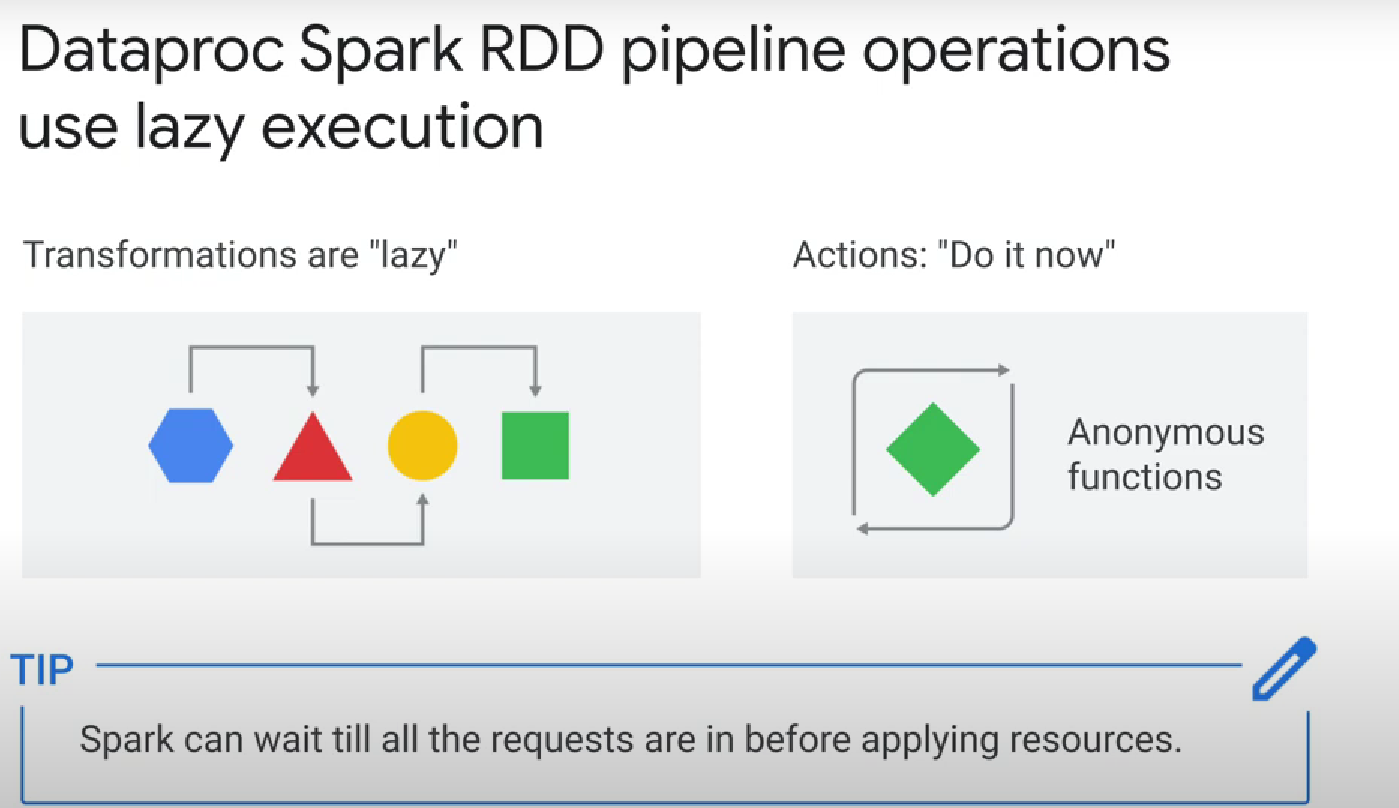


Spark:

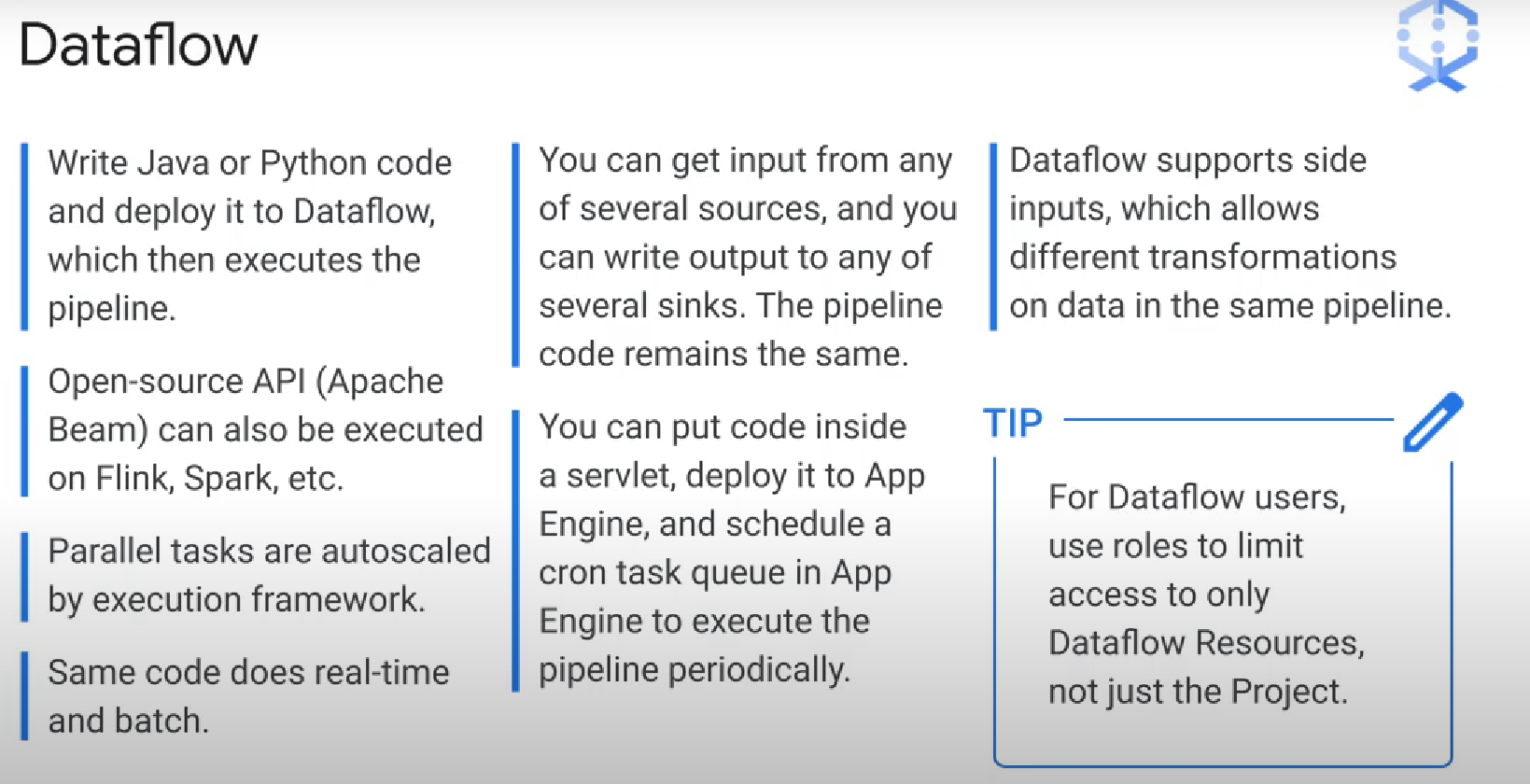
with a spark pipeline you have two different kinds of operations **transforms** and **actions**

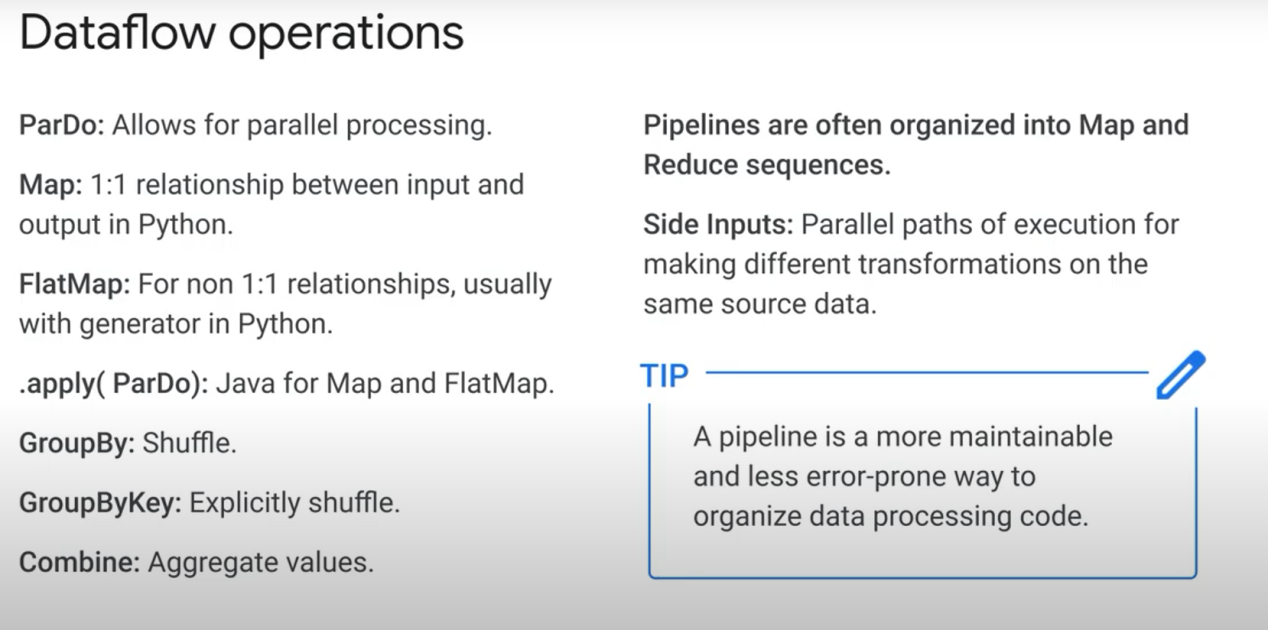


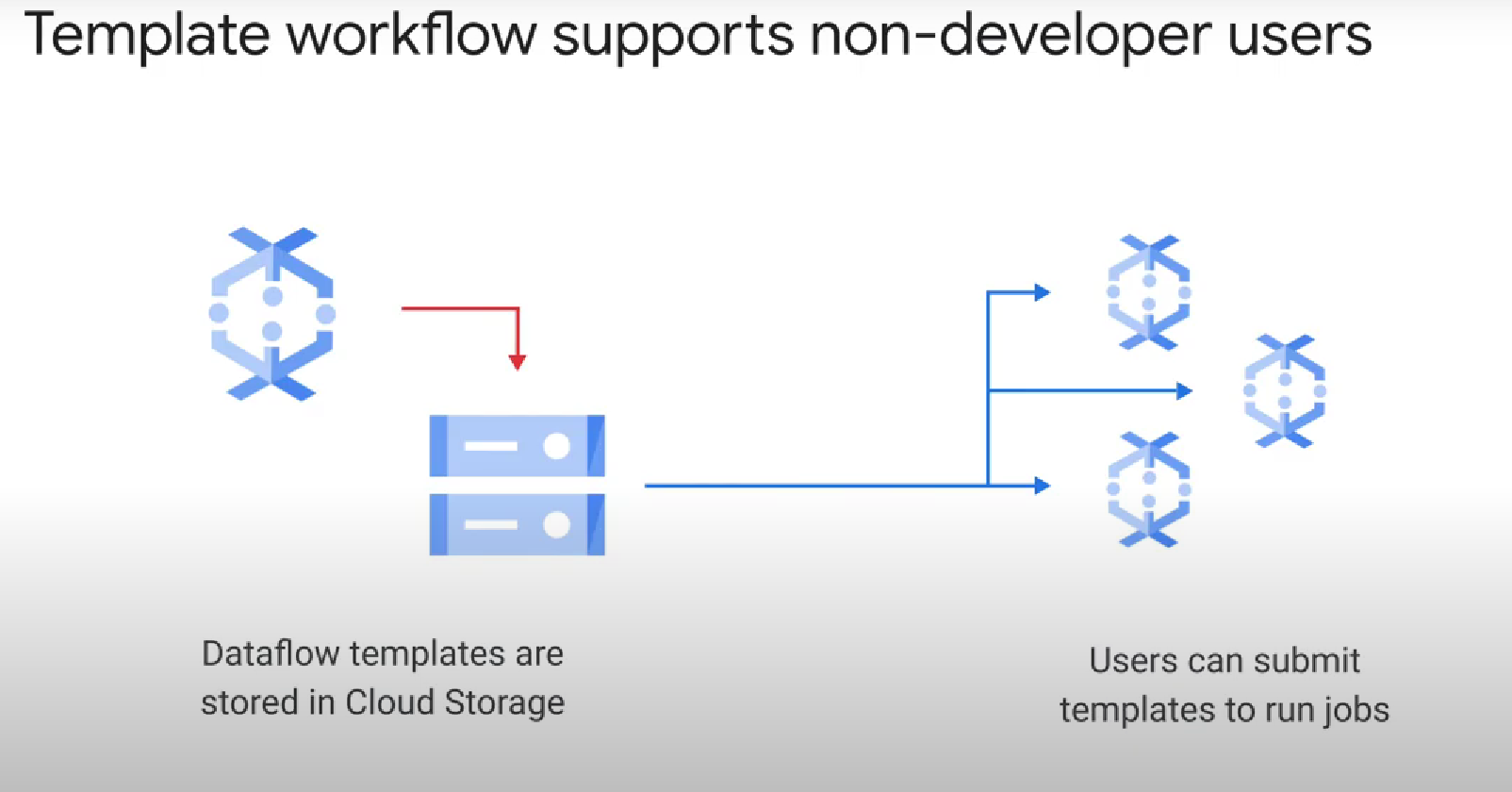
Alternative to hdfs = cloud storage

Alternative to HBase = BigTable

Note: installing and running **hadoop open source software** on **cloud dataproc** **clusters**: **use initialization actions** which are init scripts to load install and customize software







dataflow templates open up new options for separation of work (developers create the templates, users run them in jobs), and that means better security and resource accountability

Note: you can combine **Pub/Sub** with a dataflow pipeline to get **exactly once ordered processing** cloud pub sub handles exactly once delivery and cloud data flow handles deduplication ordering and windowing

**=> Separation of duties enables a scalable solution that surpasses bottlenecks in competing messaging systems**

**---------------------------------------------------------------------------------------------------------------------------------**

