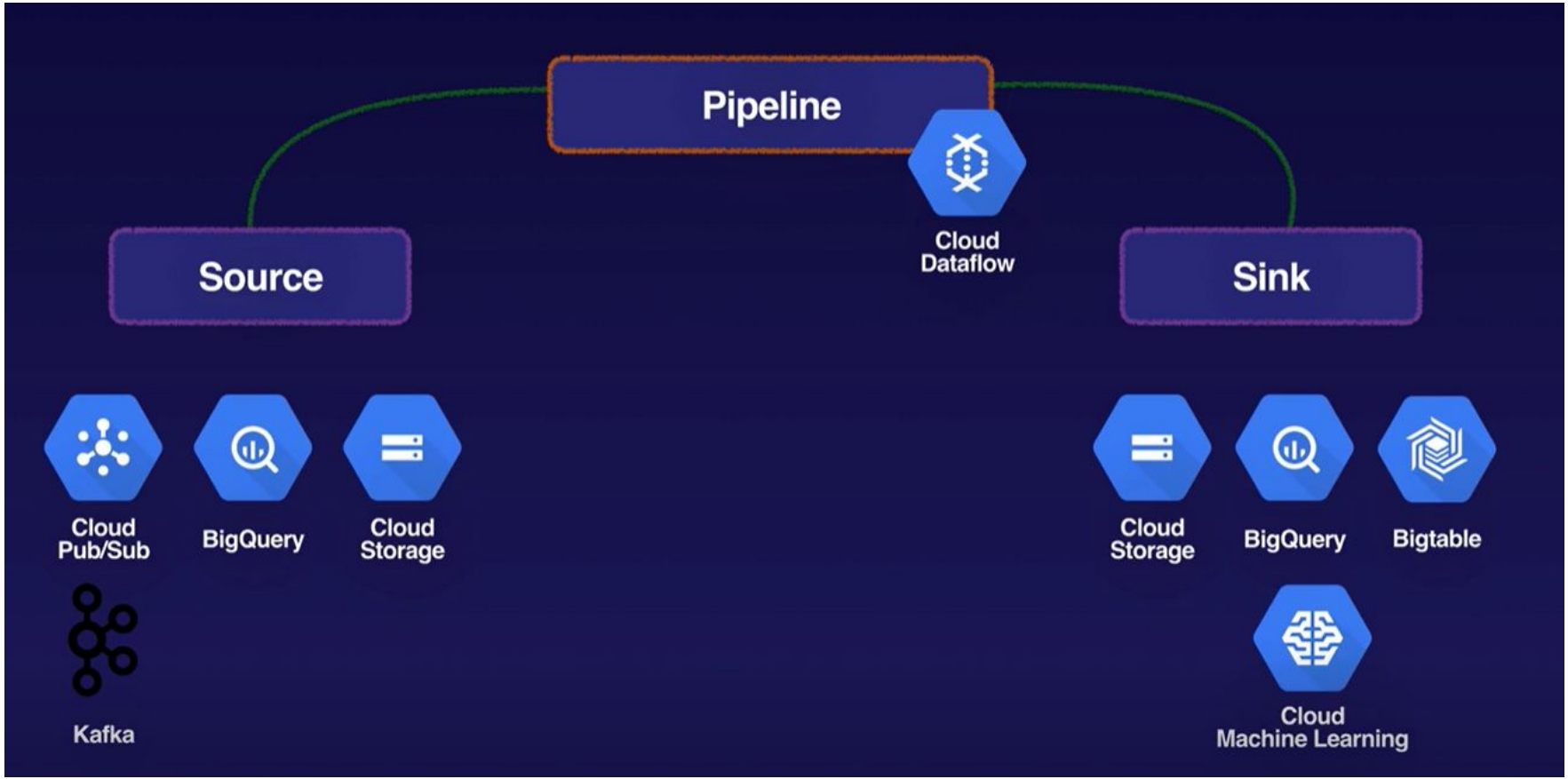
GCP Dataflow

Dataflow introduction

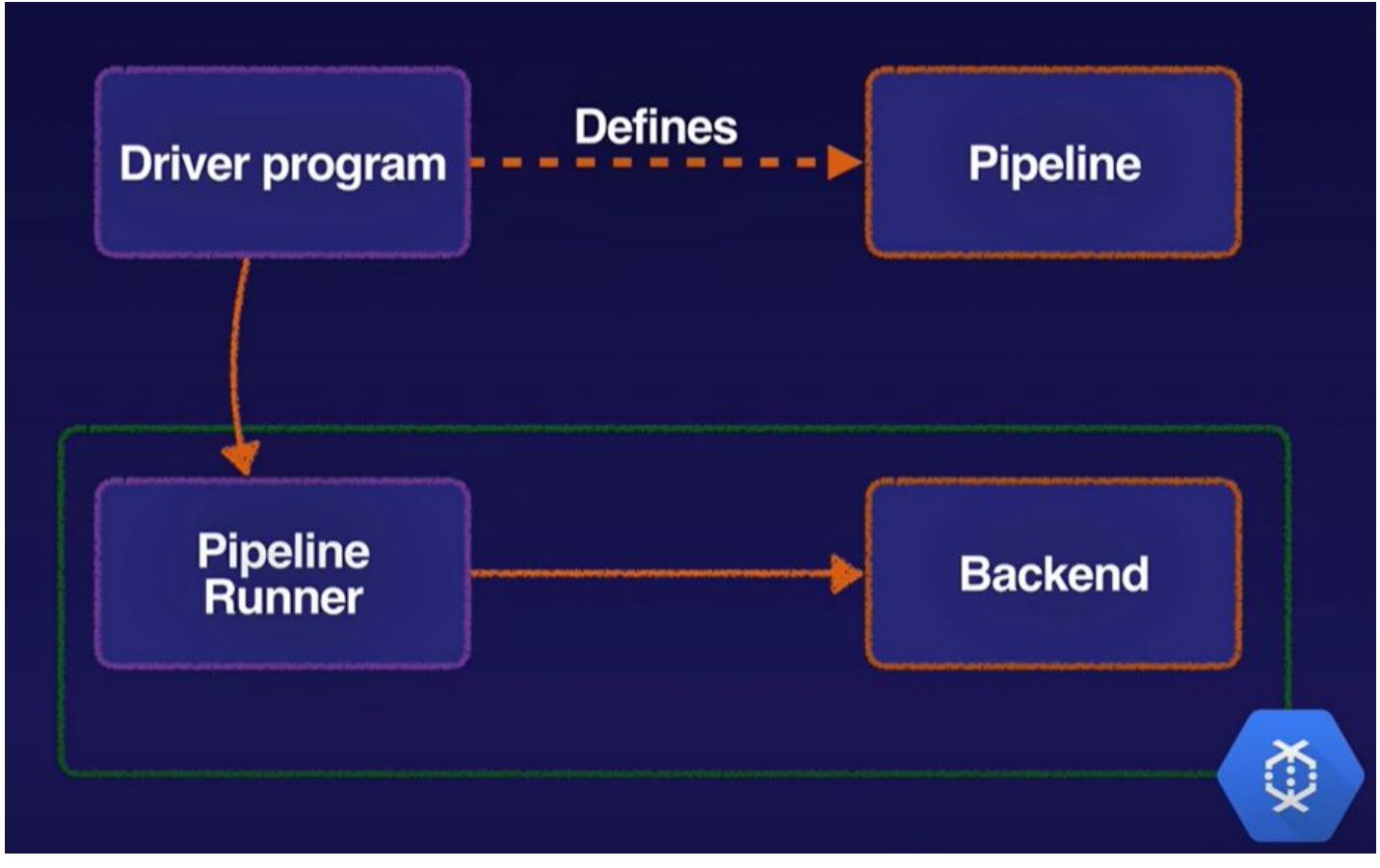
Cloud Dataflow

- Fully managed, serverless tool.
- Uses open source Apache Beam SDK.
- Supports expressive SQL, Java, and Python APIs.
- Real-time and batch processing.
- Stackdriver integration.

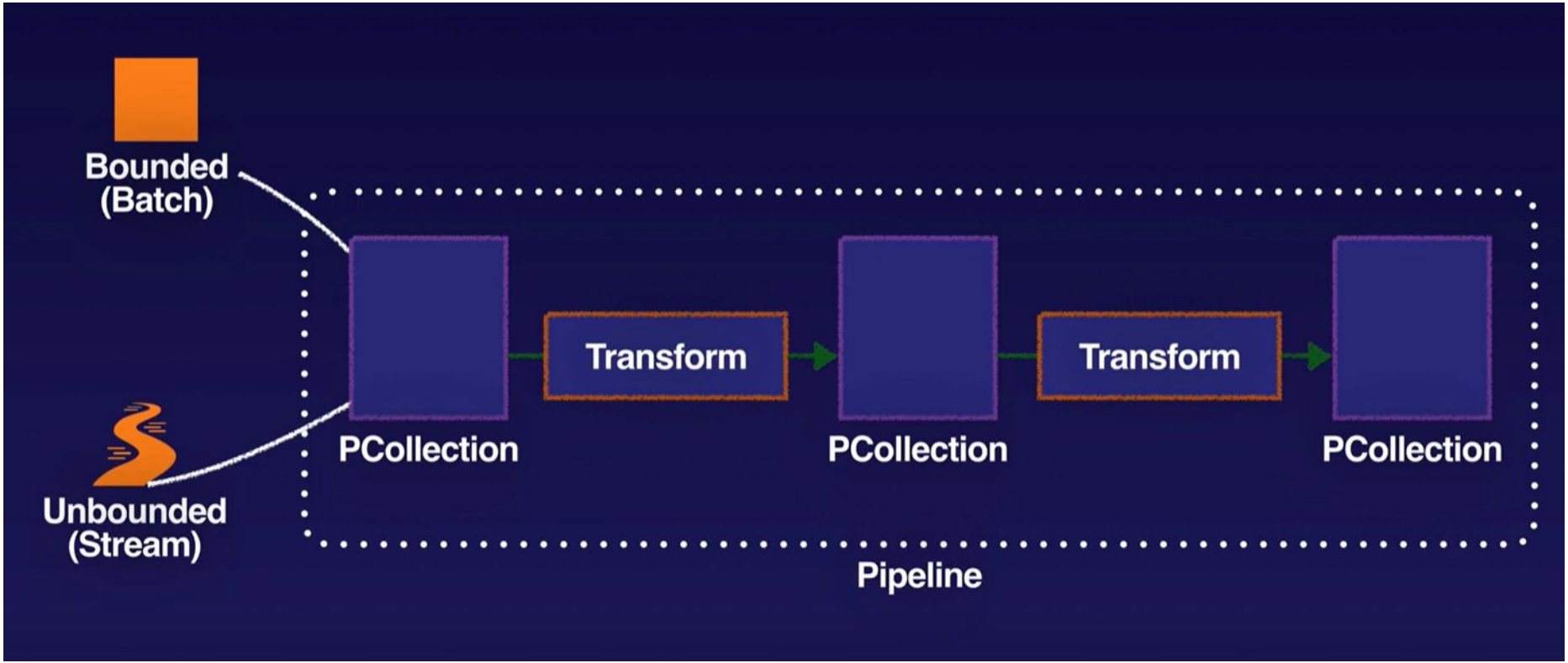
Pipeline Source and Sink



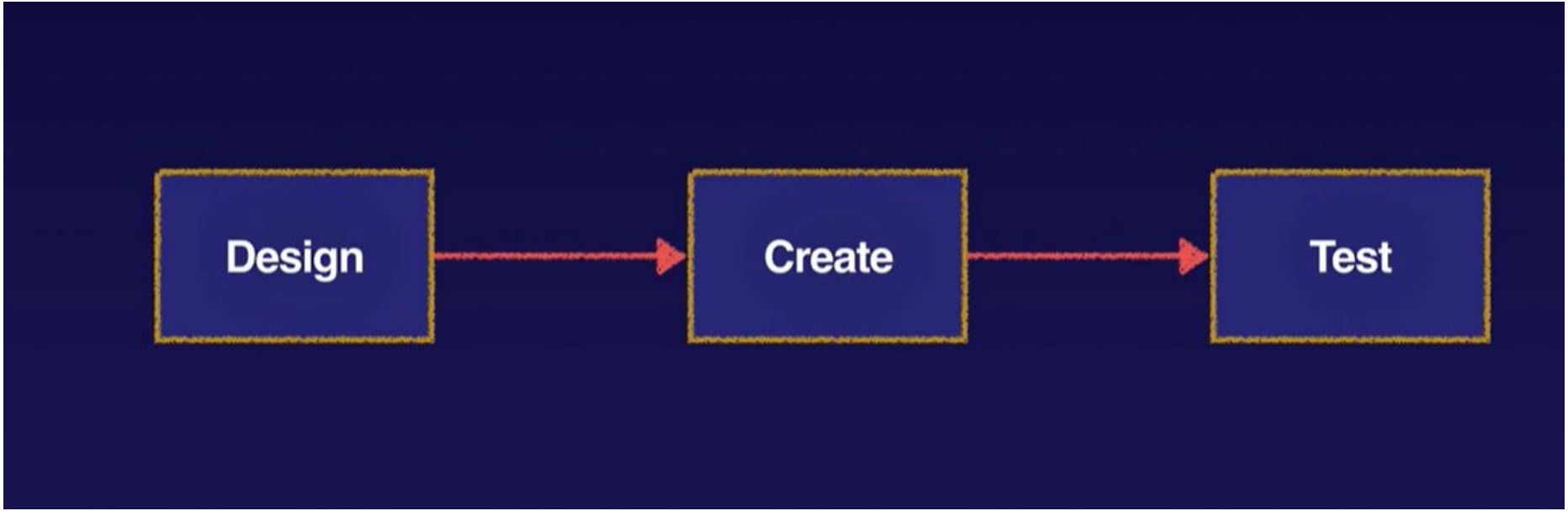
Driver Program and Runner



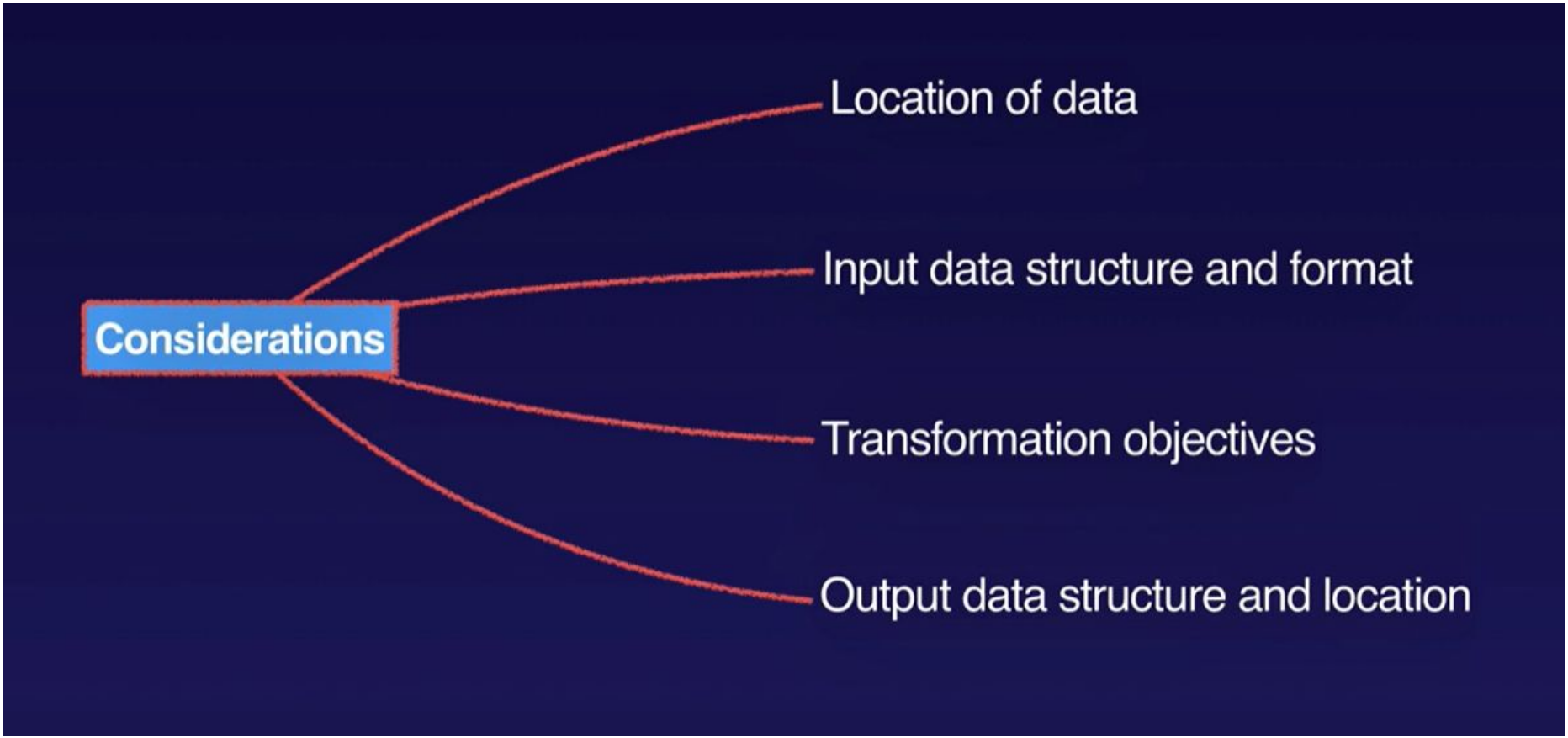
PCollections and Transforms



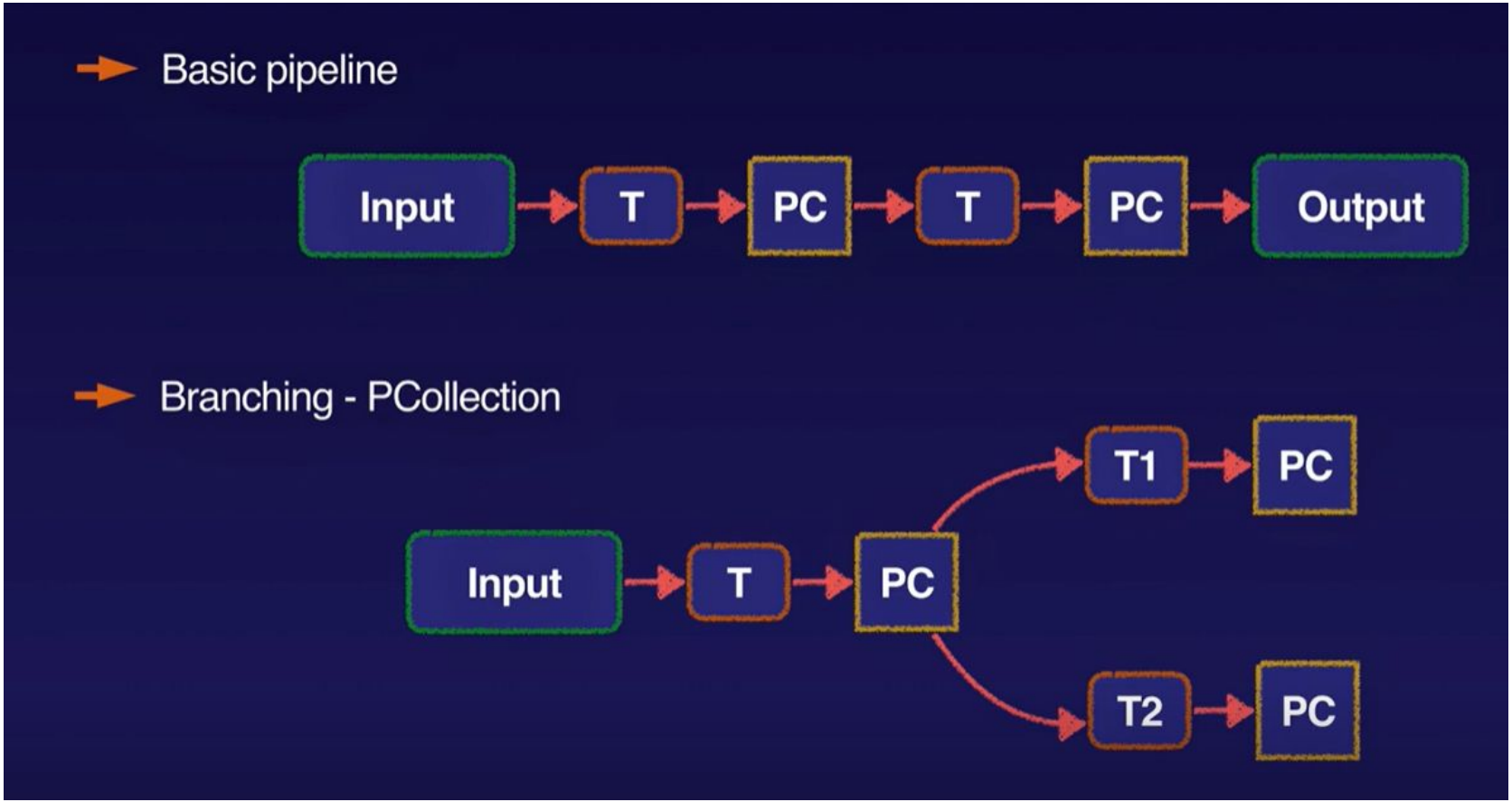
Pipeline Development Lifecycle



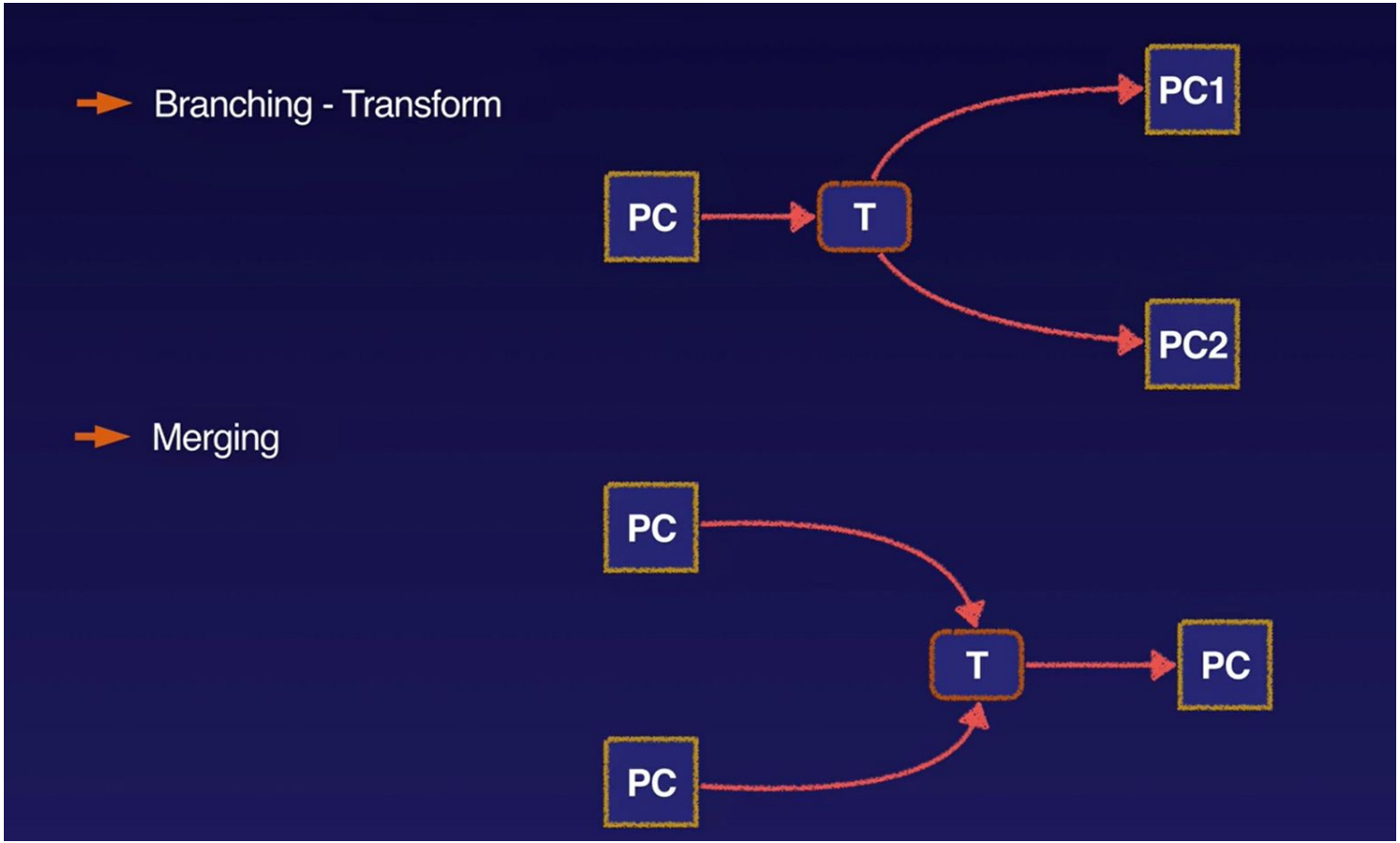
Pipeline Design Considerations



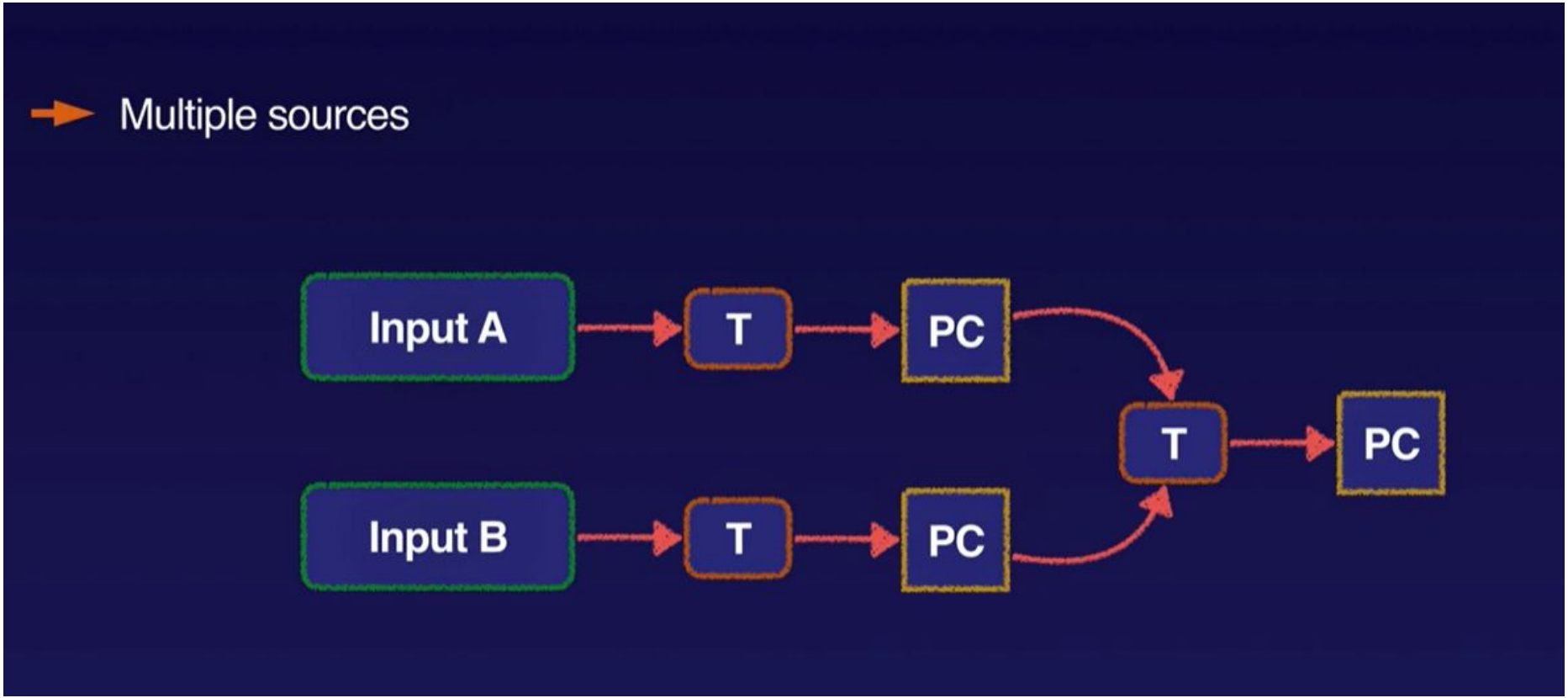
Pipeline Structure - Basic and Branching



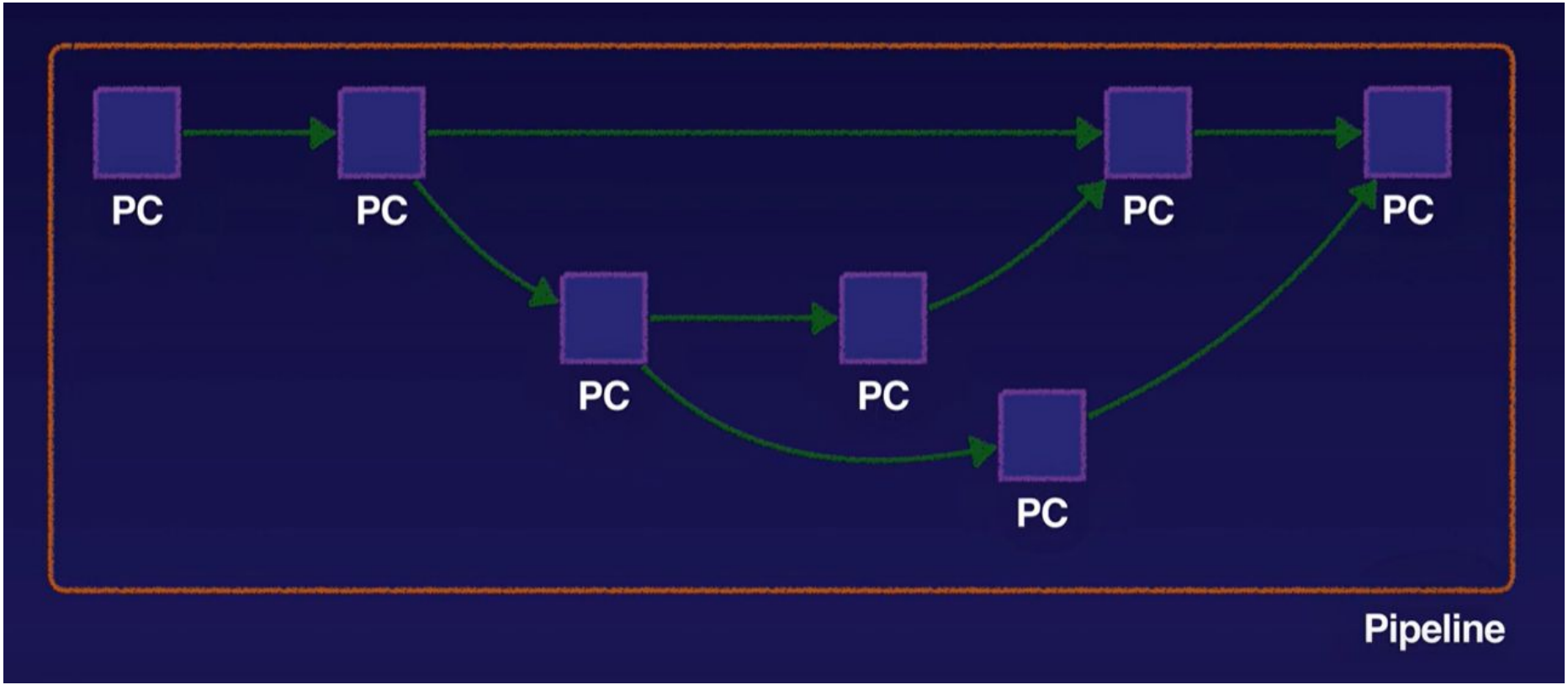
Pipeline Structure - Branching and Merging



Pipeline Structure - Branching and Merging



Pipeline Graph - Directed Acyclic Graph



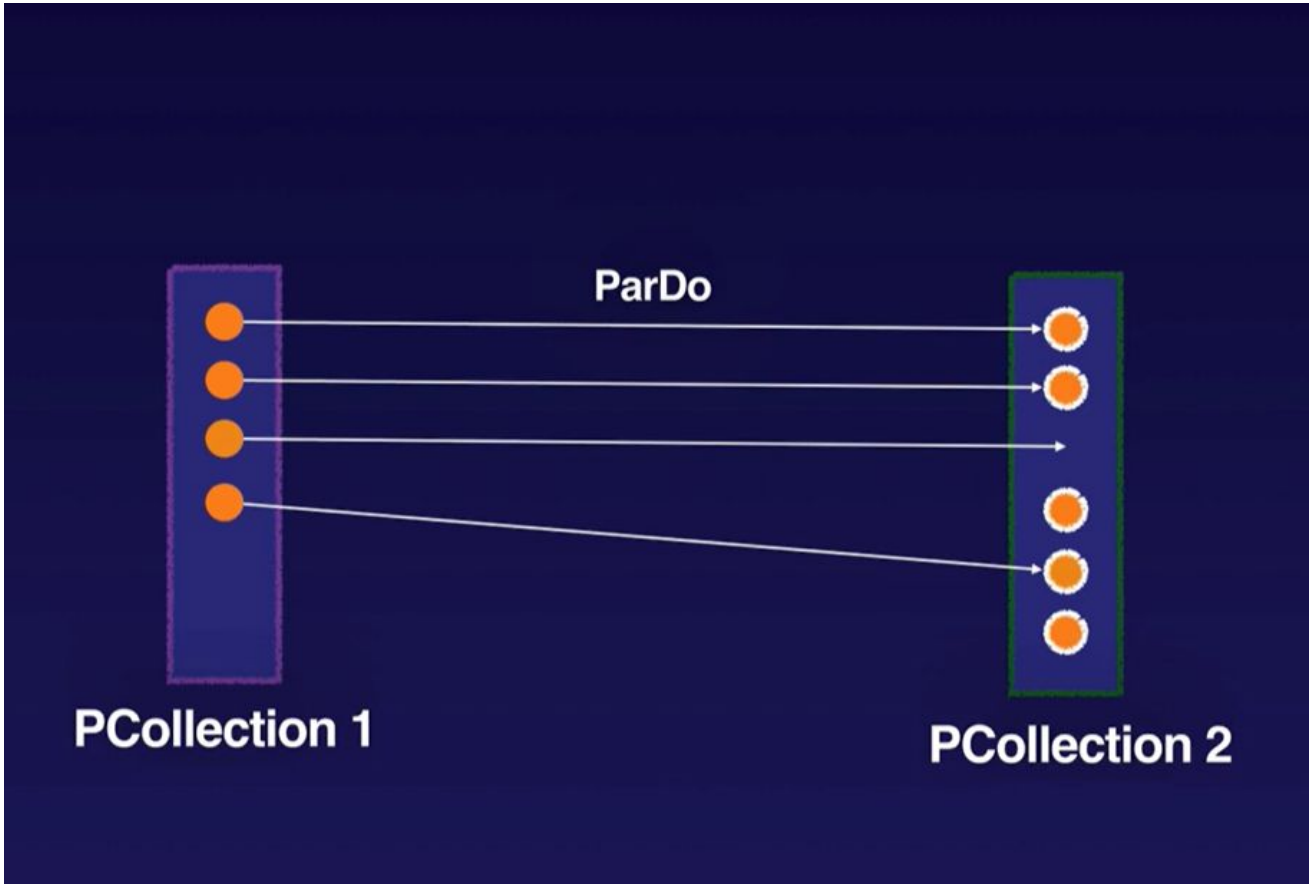
Pipeline Creation

Driver Program (design time)

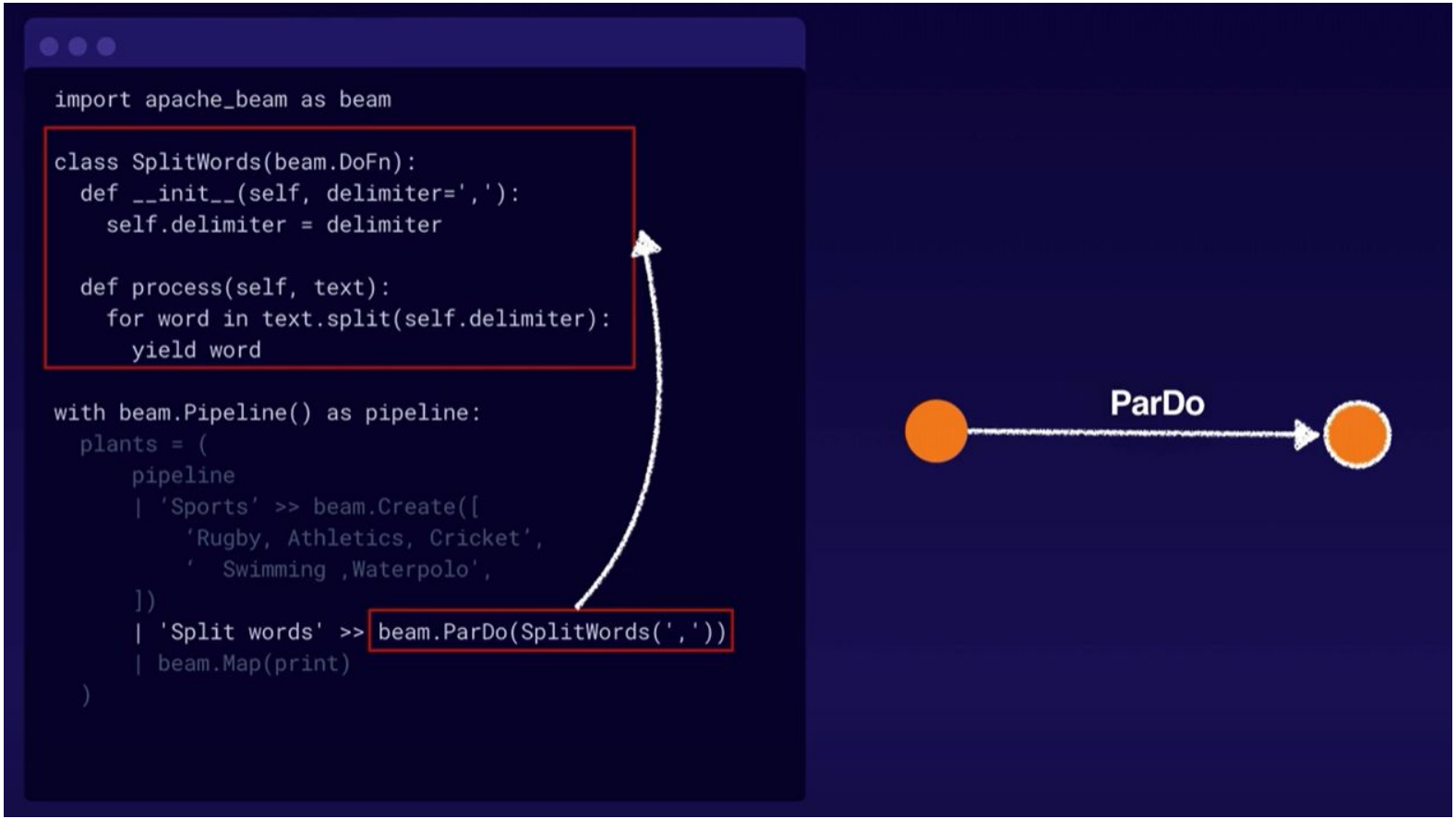
- Create pipeline object
- Create a PCollection using read or create transform
- Apply multiple transforms as required
- Write out final PCollection

Dataflow Pipeline Concepts

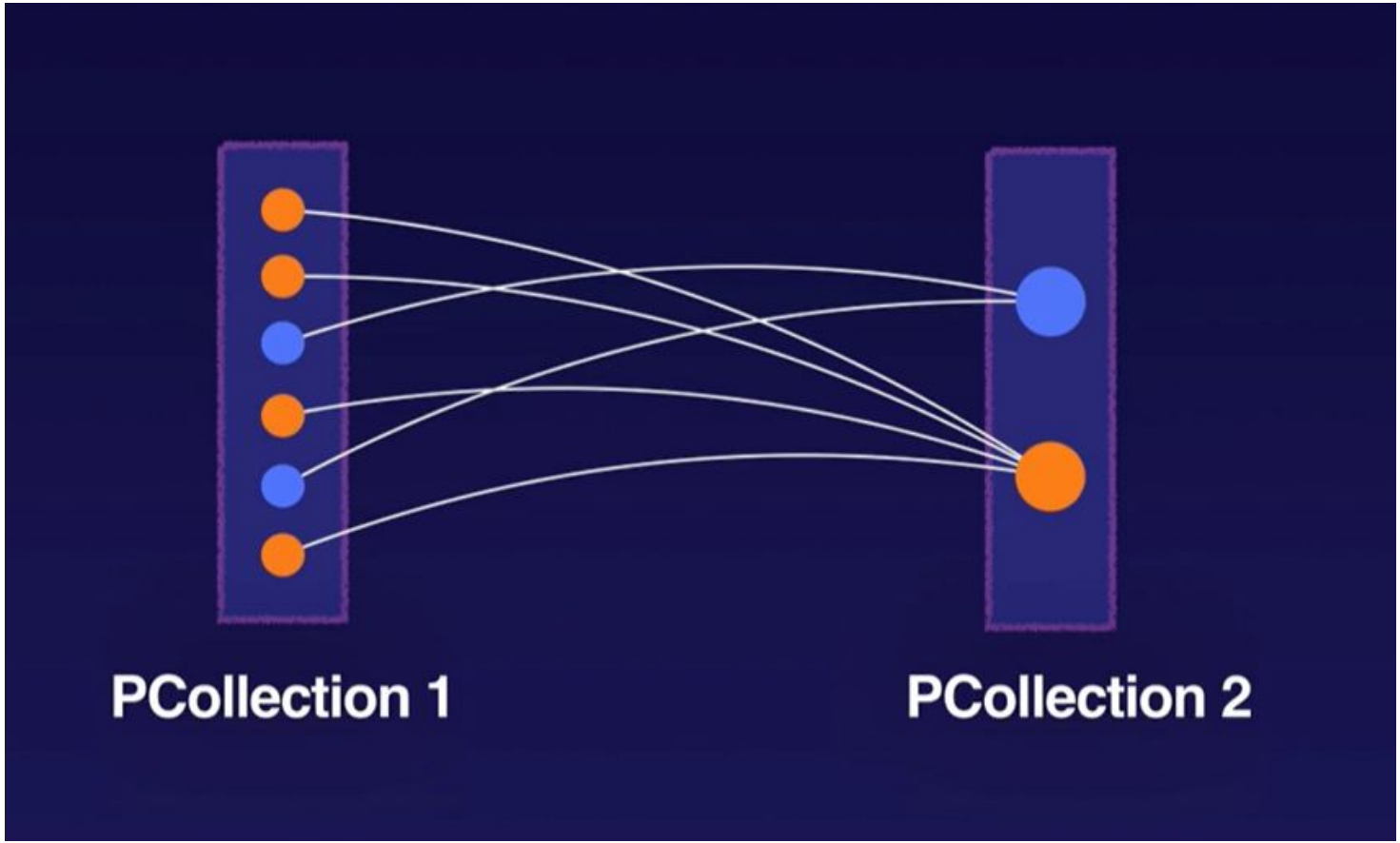
ParDo



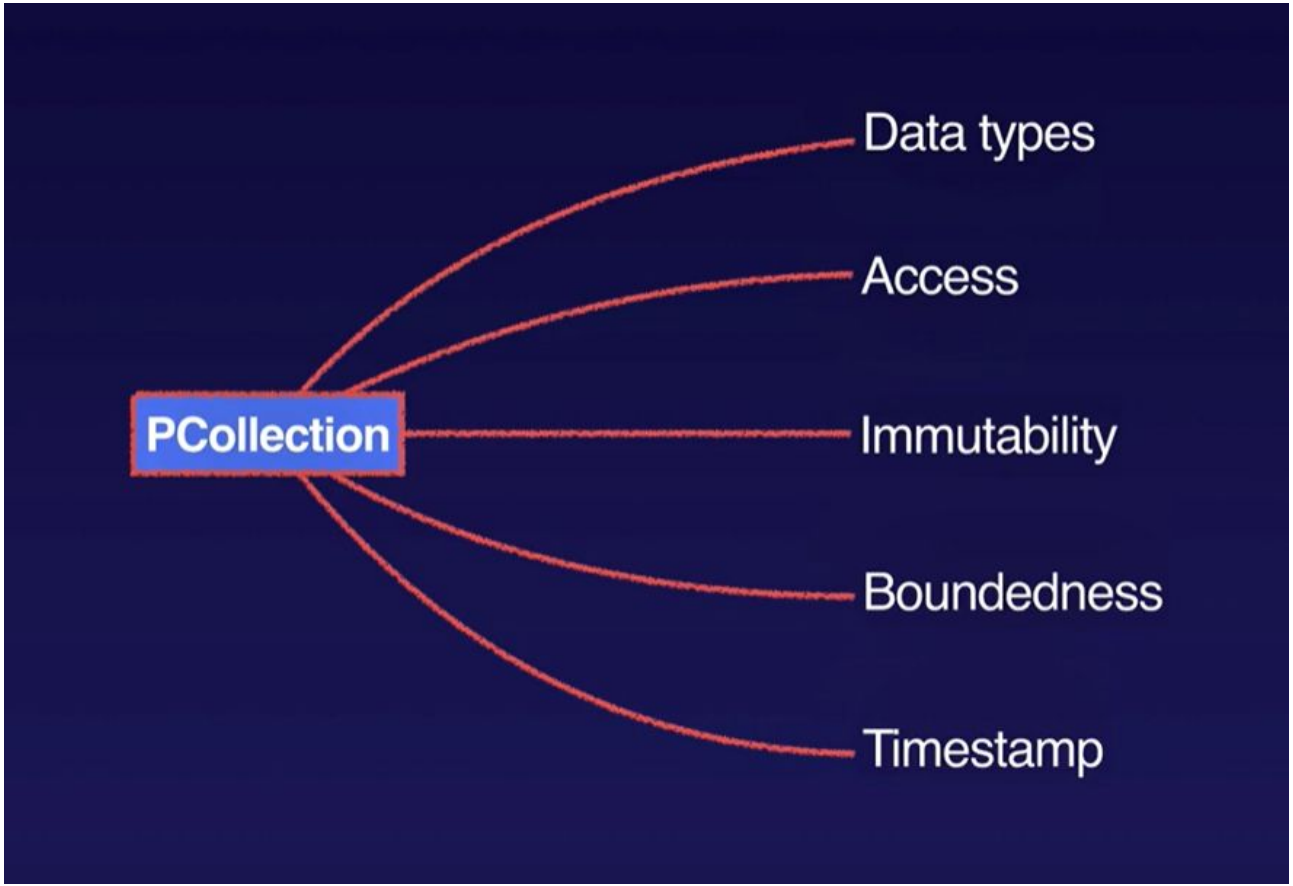
User-defined function (UDF)



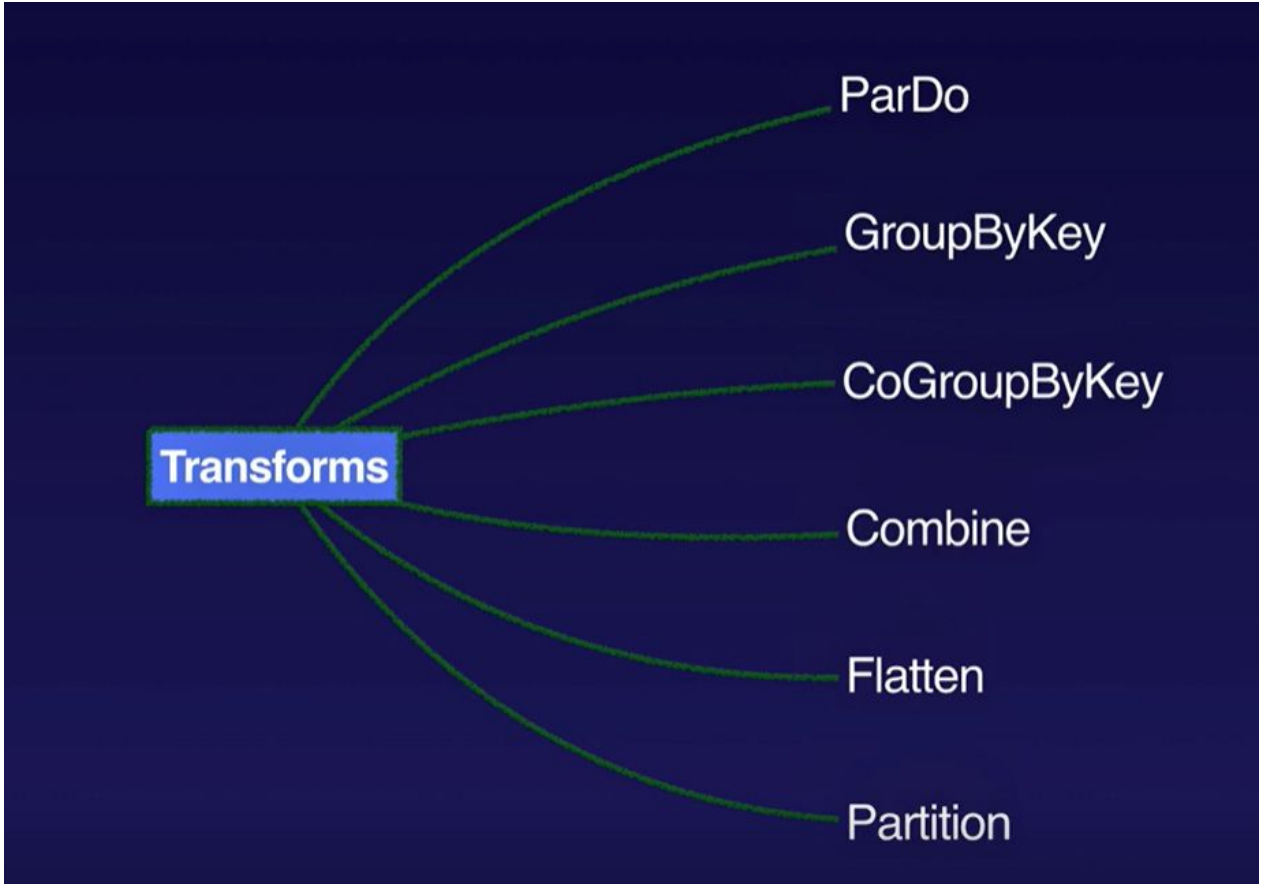
Aggregation



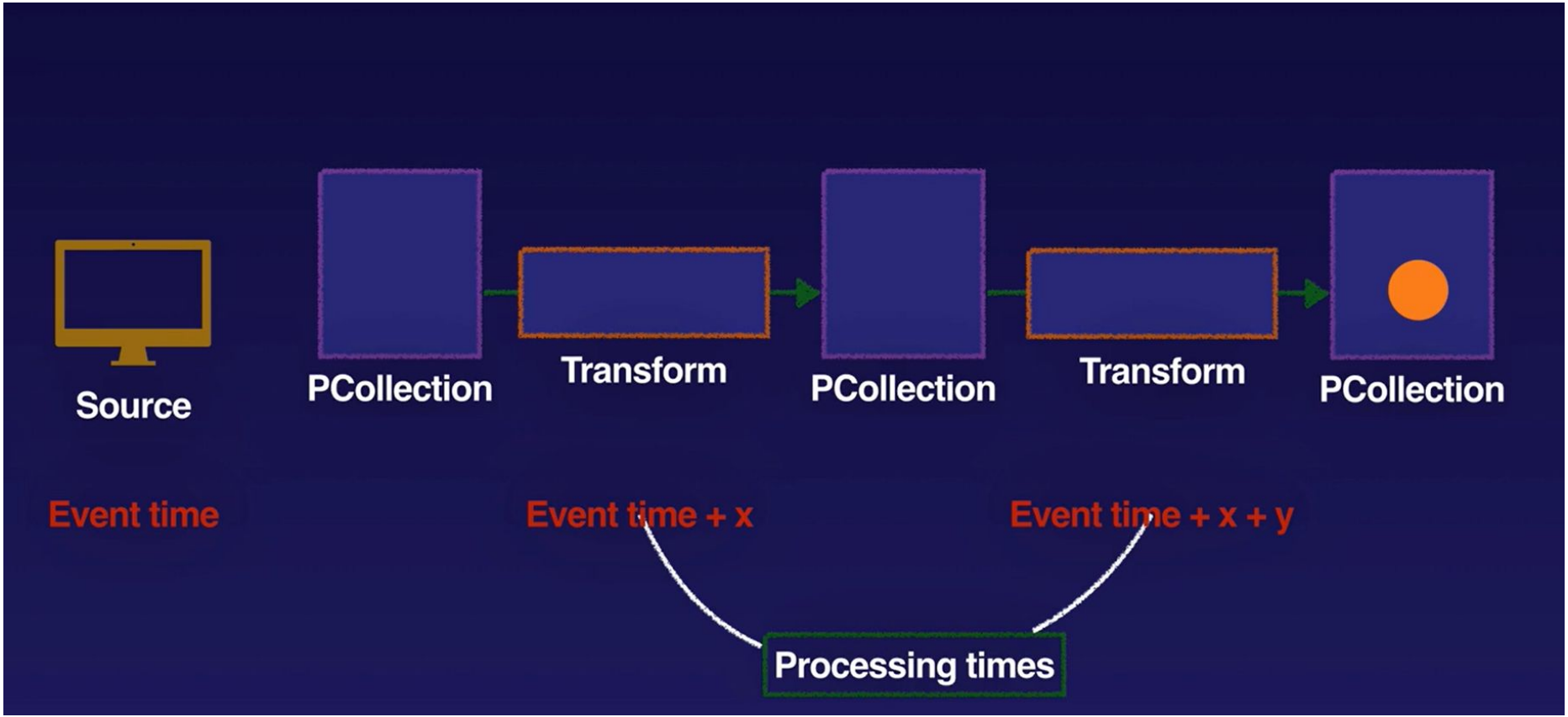
Characteristics of PCollections



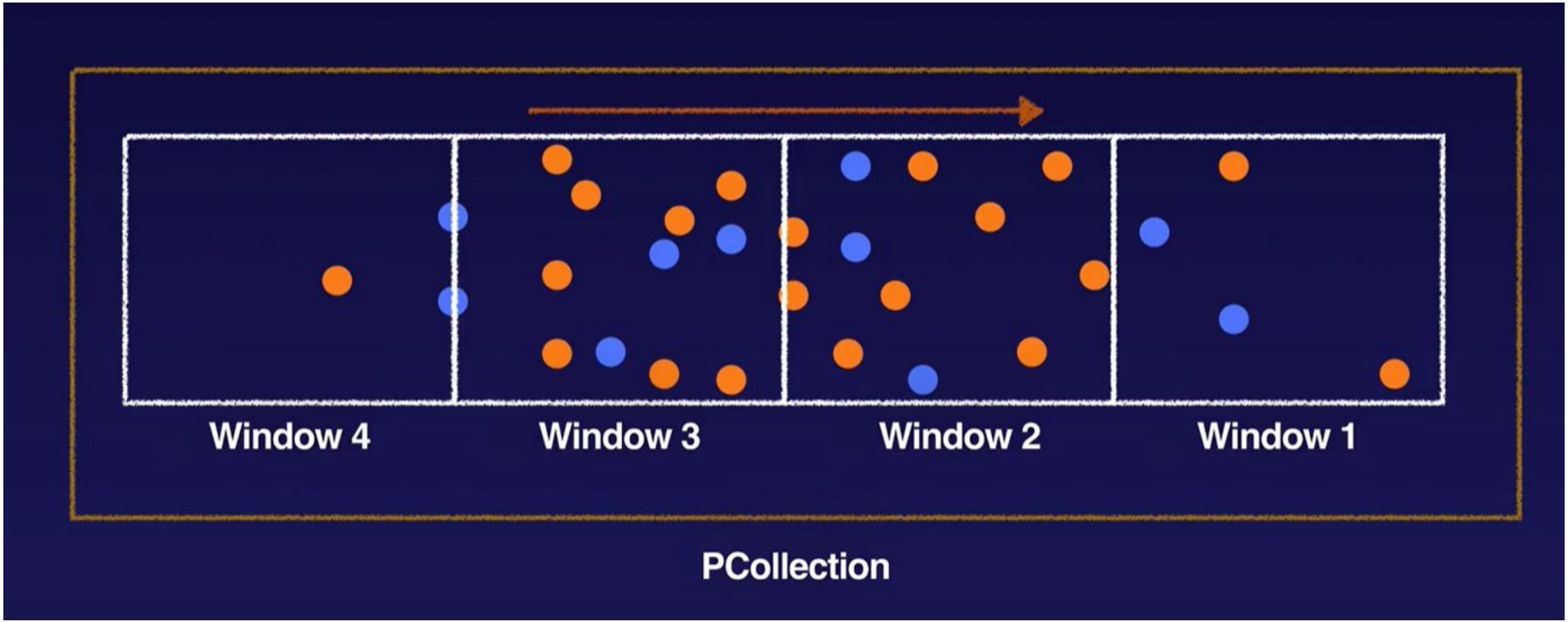
Core Beam transforms



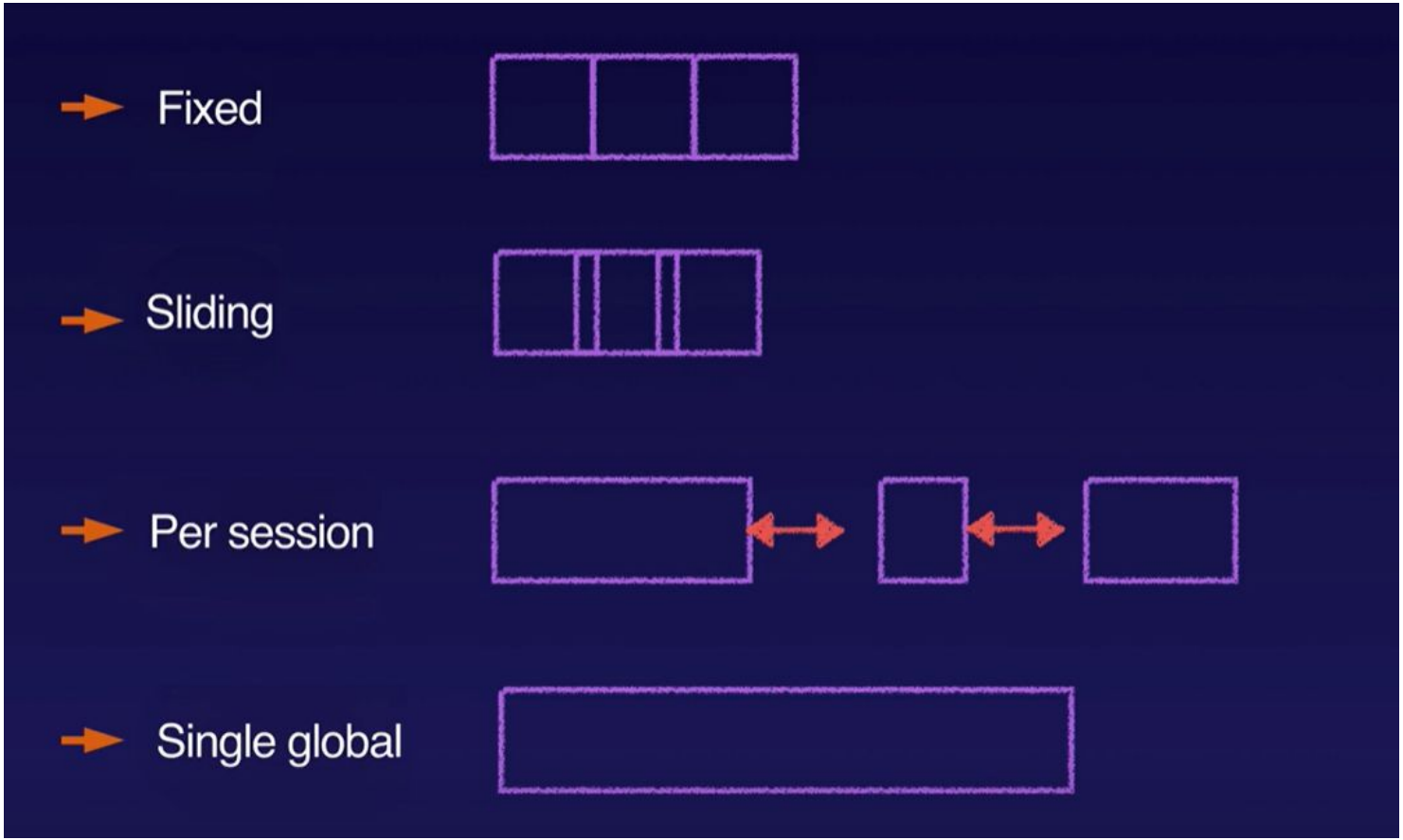
Advanced Dataflow Concepts - Event time



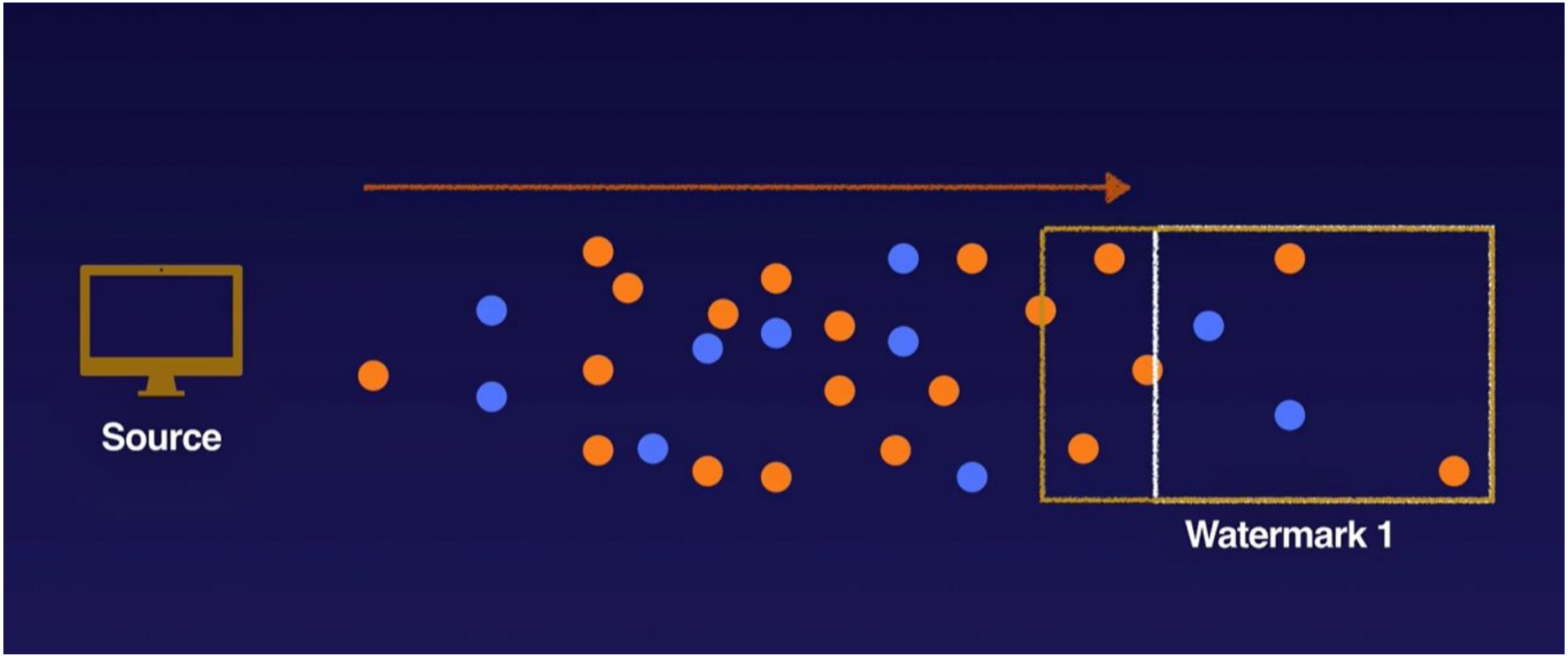
Advanced Dataflow Concepts - Windowing



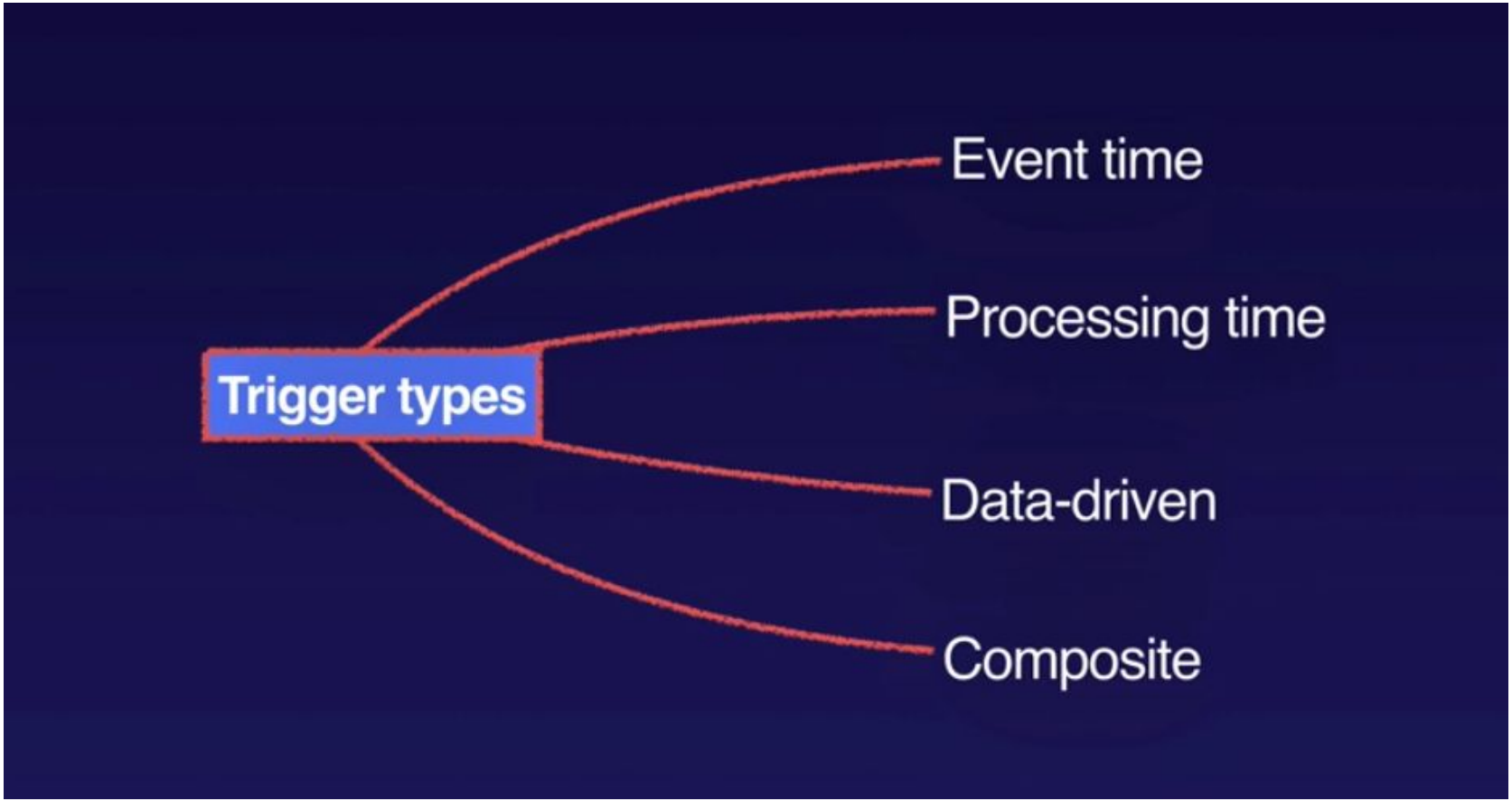
Advanced Dataflow Concepts



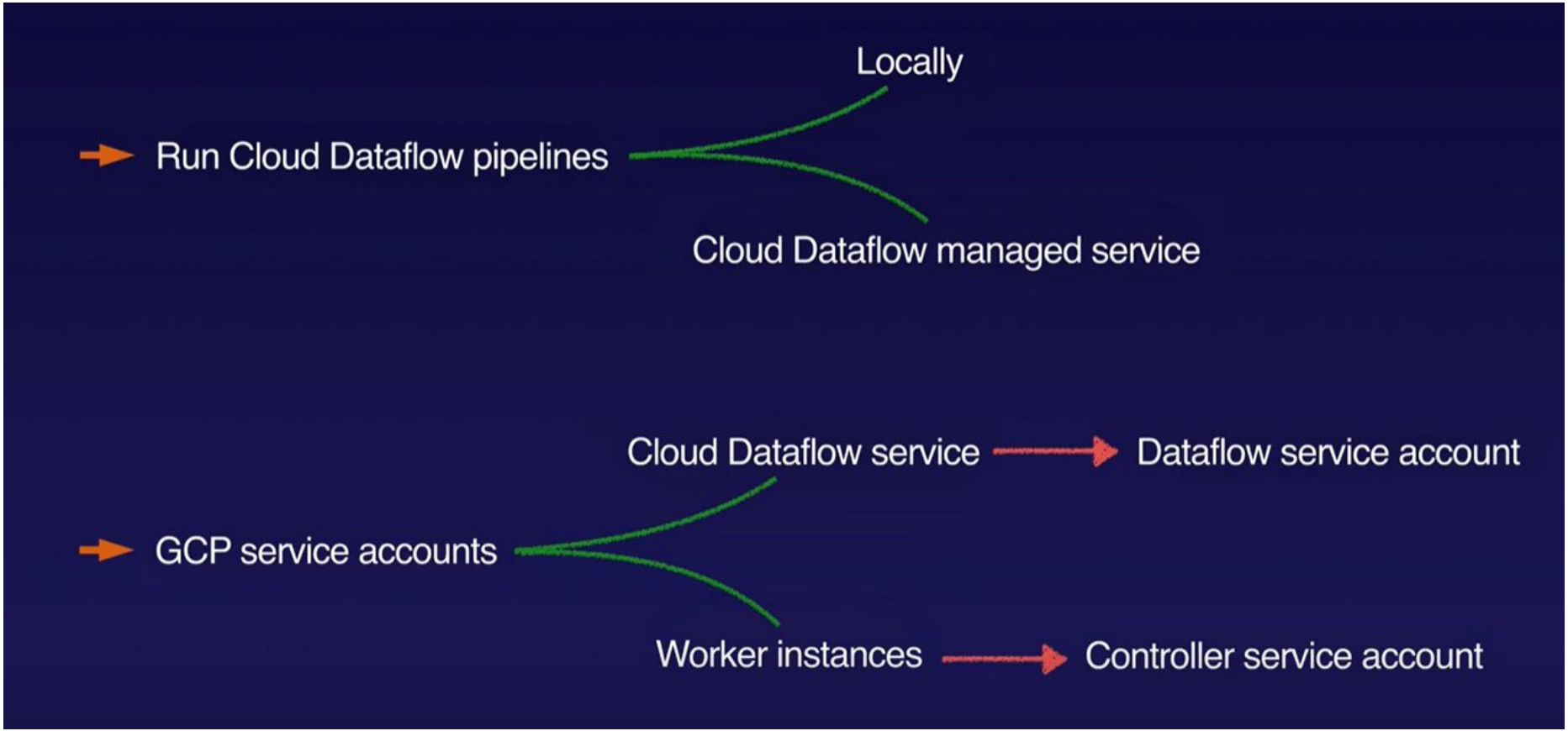
Watermarks



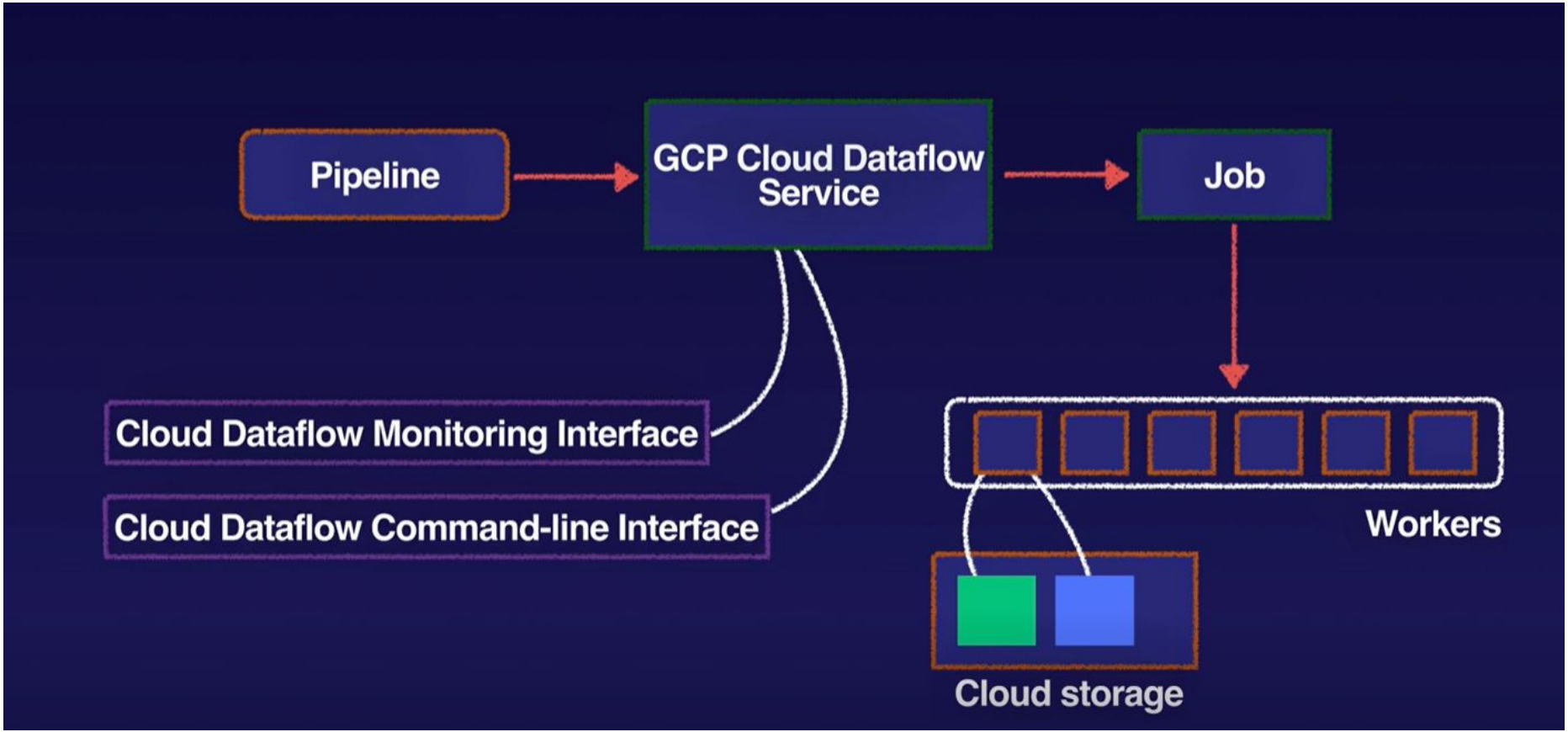
Triggers


Trigger types
Event time
Processing time
Data-driven
Composite

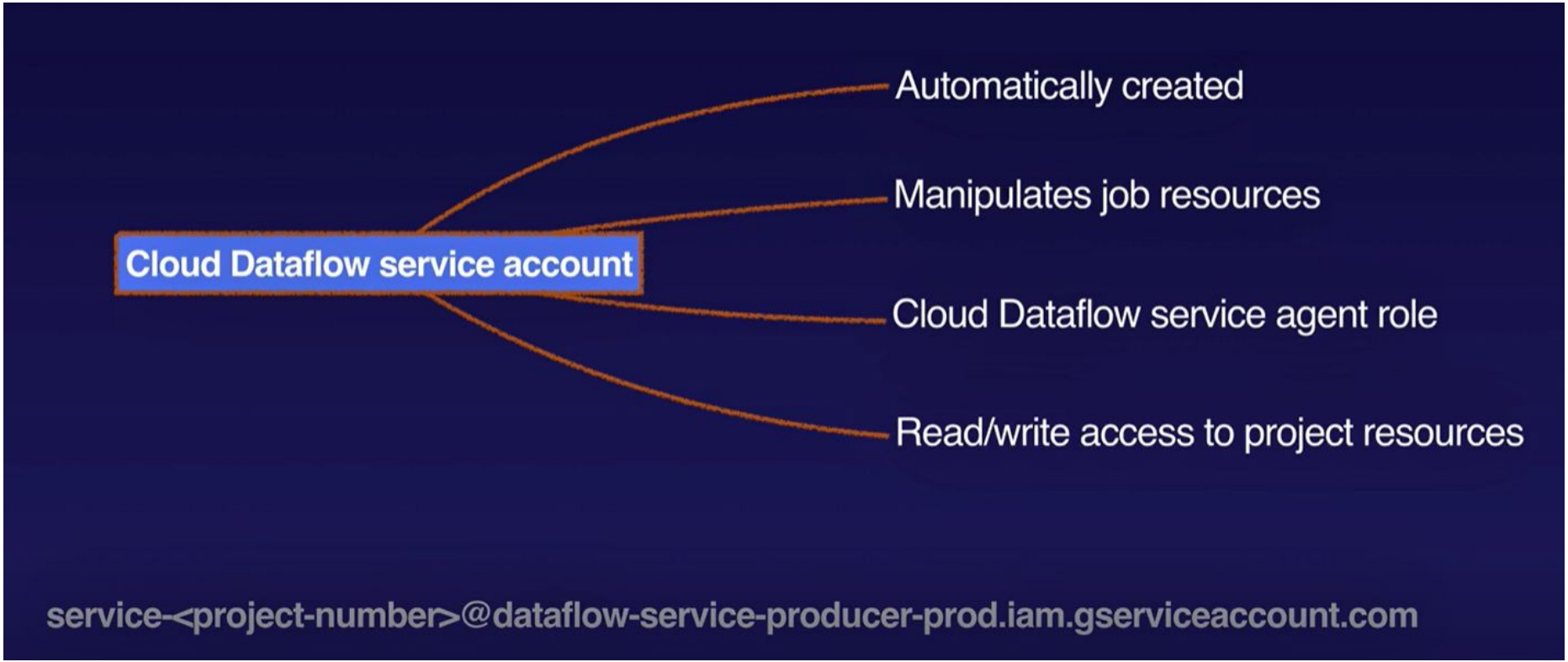
Pipeline Access



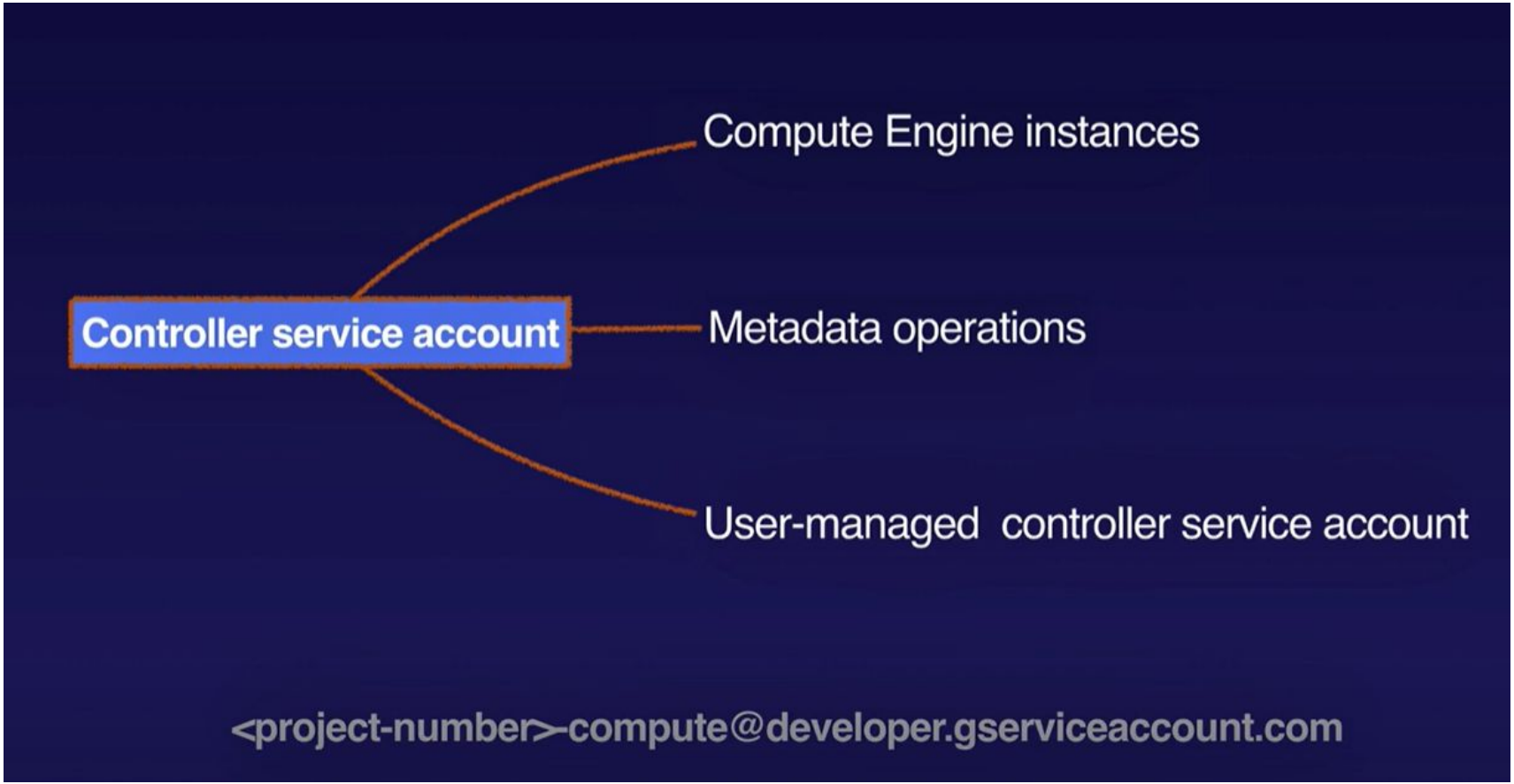
Cloud Dataflow Managed Service



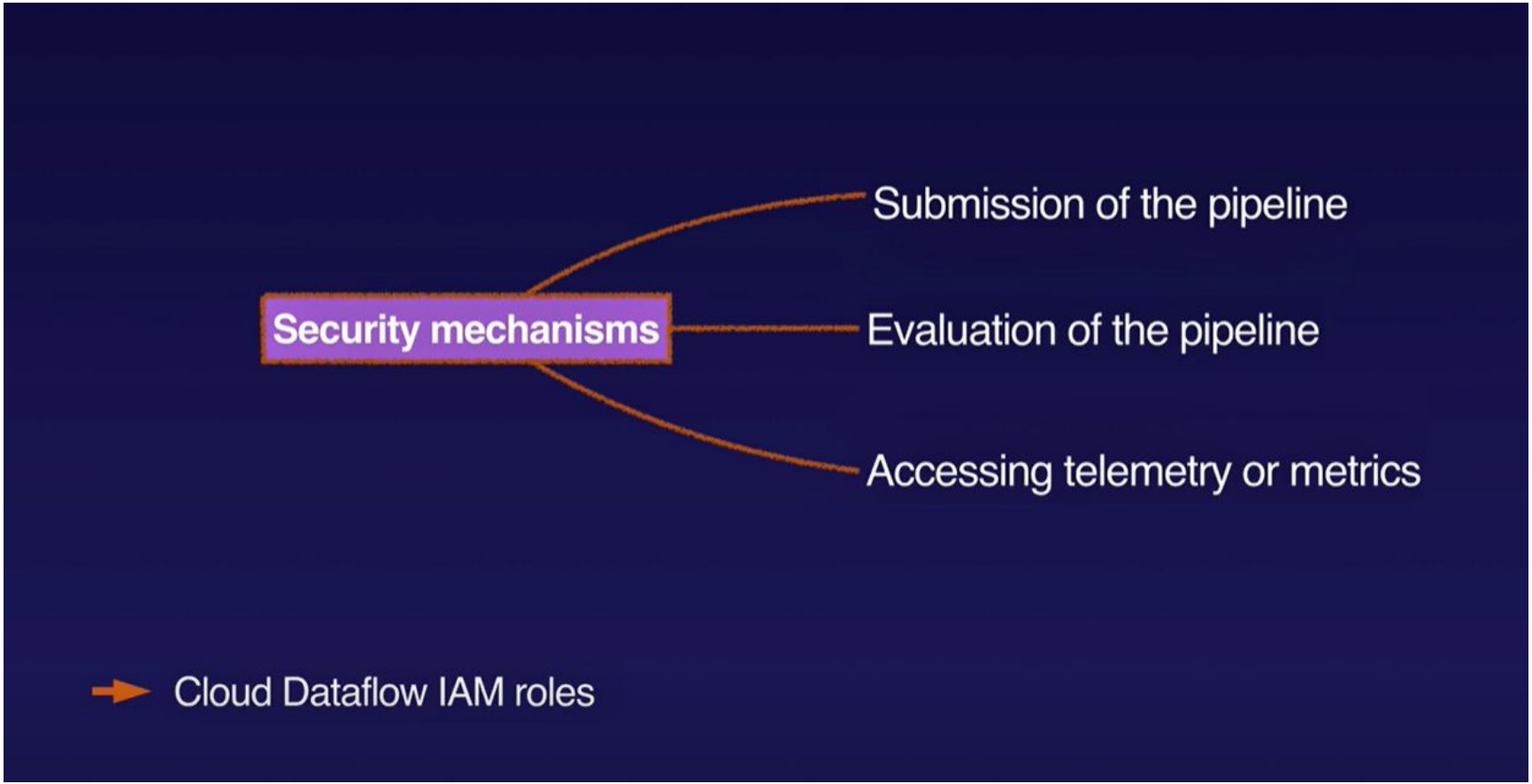
Cloud Dataflow Service Account



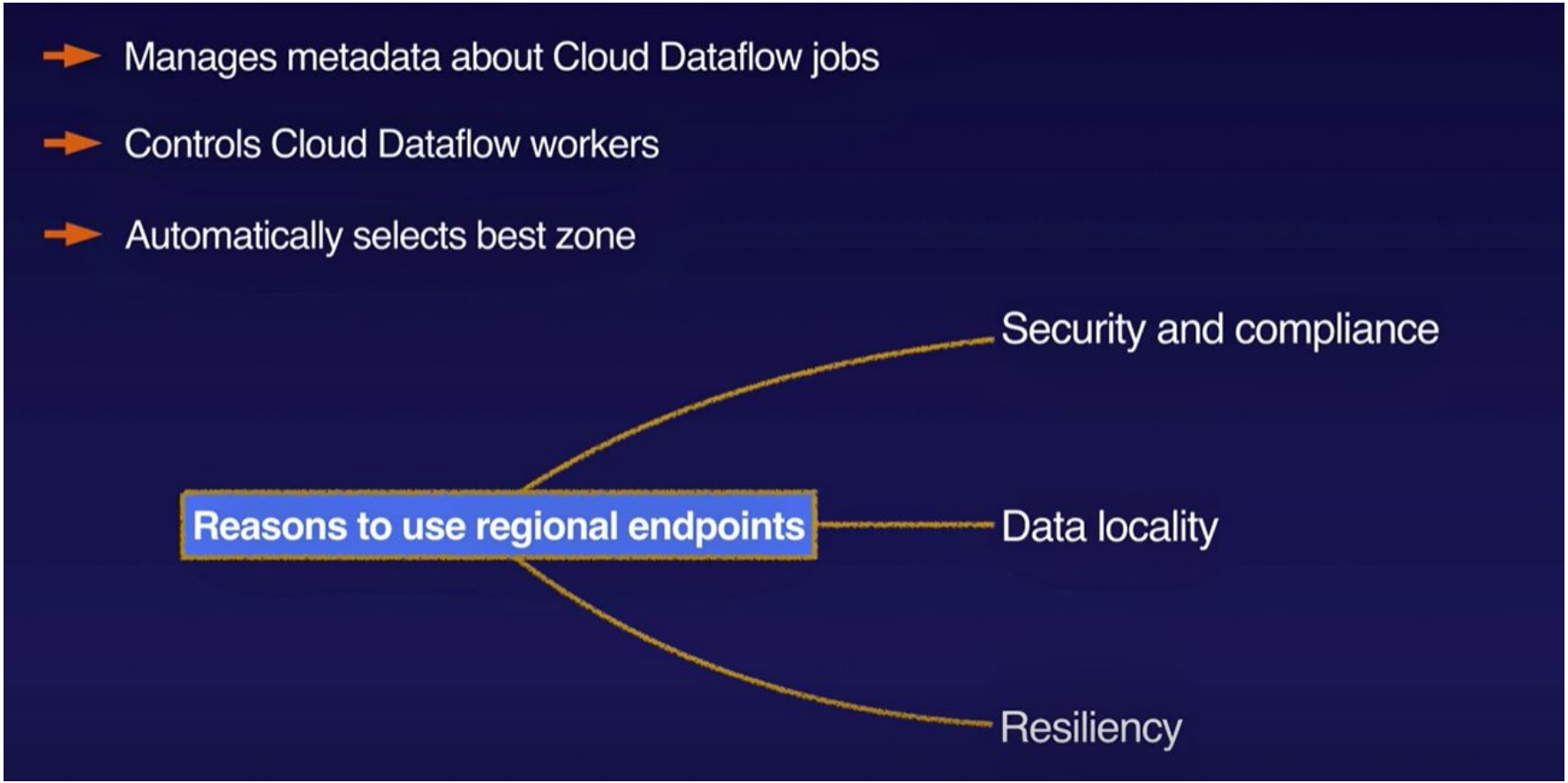
Service Accounts



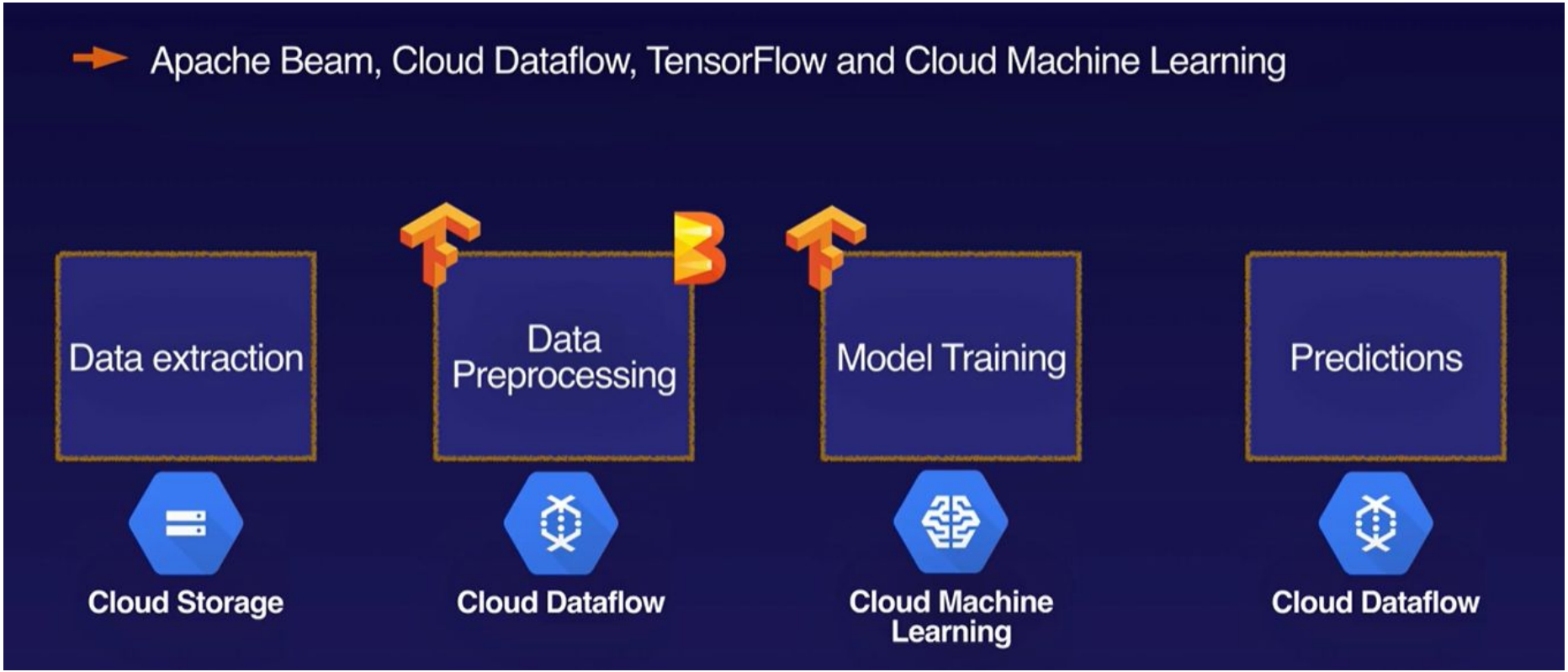
Access and Security



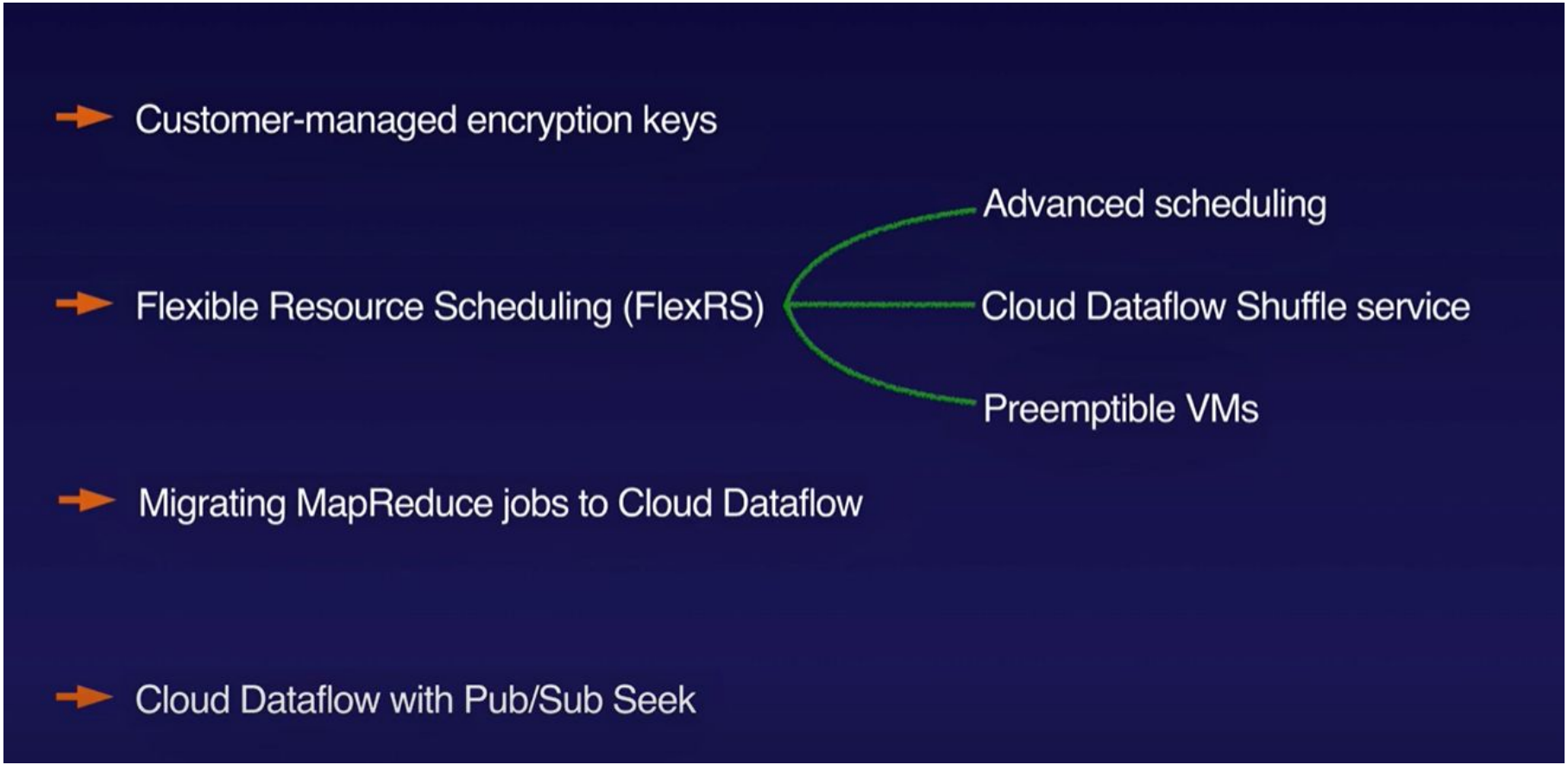
Using Dataflow - Regional Endpoints



Machine Learning with Dataflow



Using Dataflow



Cloud Dataflow SQL