Creating Pipelines from Google Templates

Janani Ravi
CO-FOUNDER, LOONYCORN

www.loonycorn.com

Overview

Templates for ease of pipeline creation

Deploying pipelines from templates

Working with streaming data using Pub/Sub

Dataflow Templates

Dataflow is very hard for non-developers to use - Dataflow Templates attempt to change that

Dataflow Templates

Dataflow can be intimidating to get started with

- Needs dependencies to be set up
- Code to be written

Dataflow Templates are a great way to get started quickly

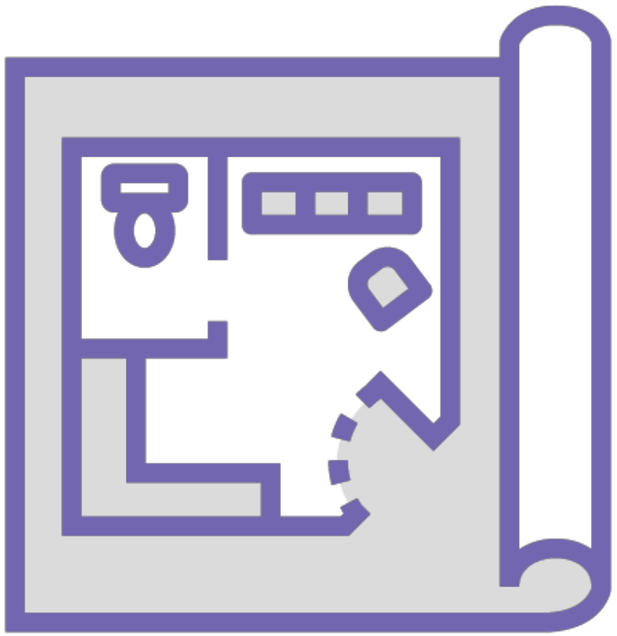
Dataflow Templates

Start with Google-provided templates for common tasks

Do not recompile code

Use runtime parameters to customize execution

Dataflow Templates

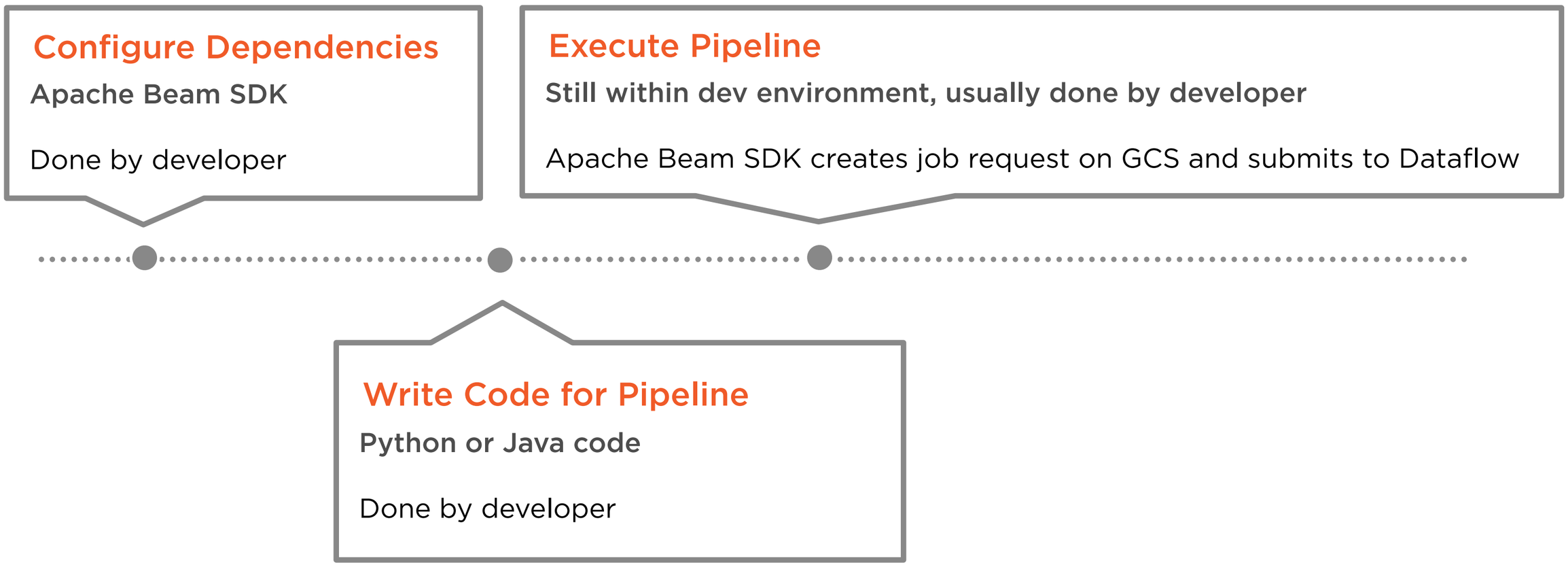


No need to configure dependencies

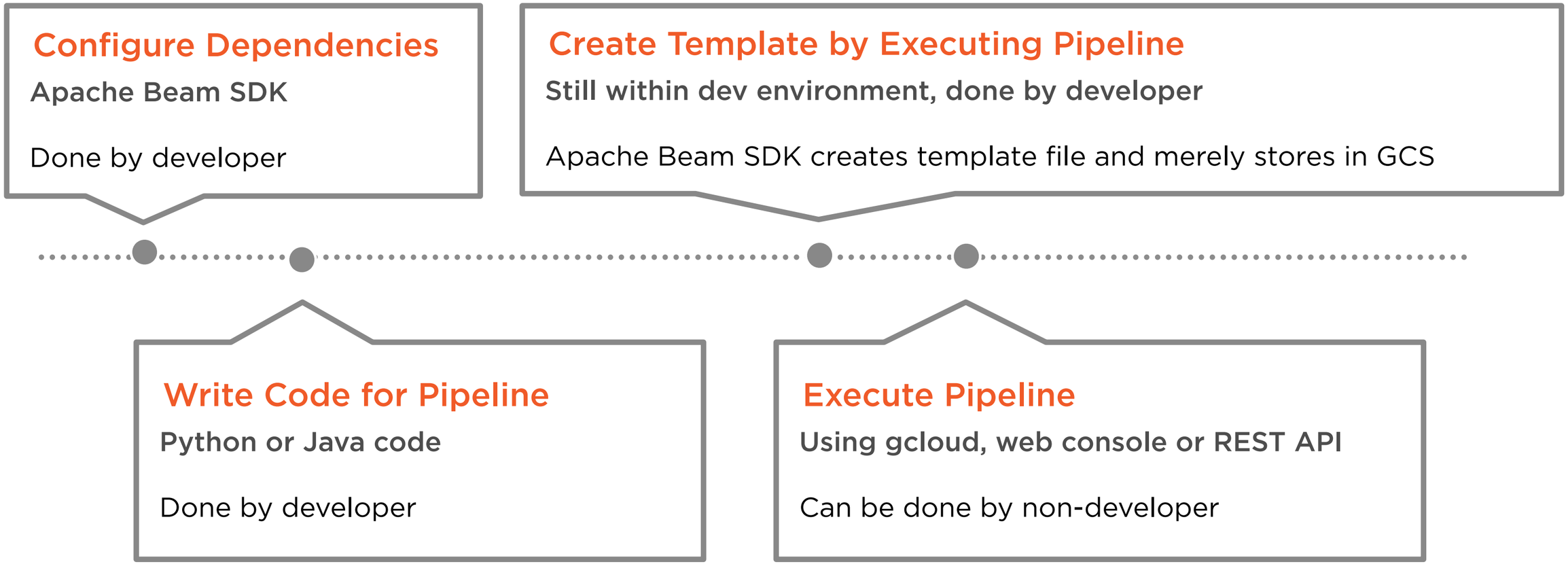
Have developers define templates

Non-developers can use those templates without writing any code

Traditional Dataflow Jobs



Templated Dataflow Jobs



Google-provided Templates

BigTable to Cloud Storage

Pub/Sub to BigQuery

Cloud Storage to Pub/Sub

...

Many more intra-GCP transfers

Pub/Sub

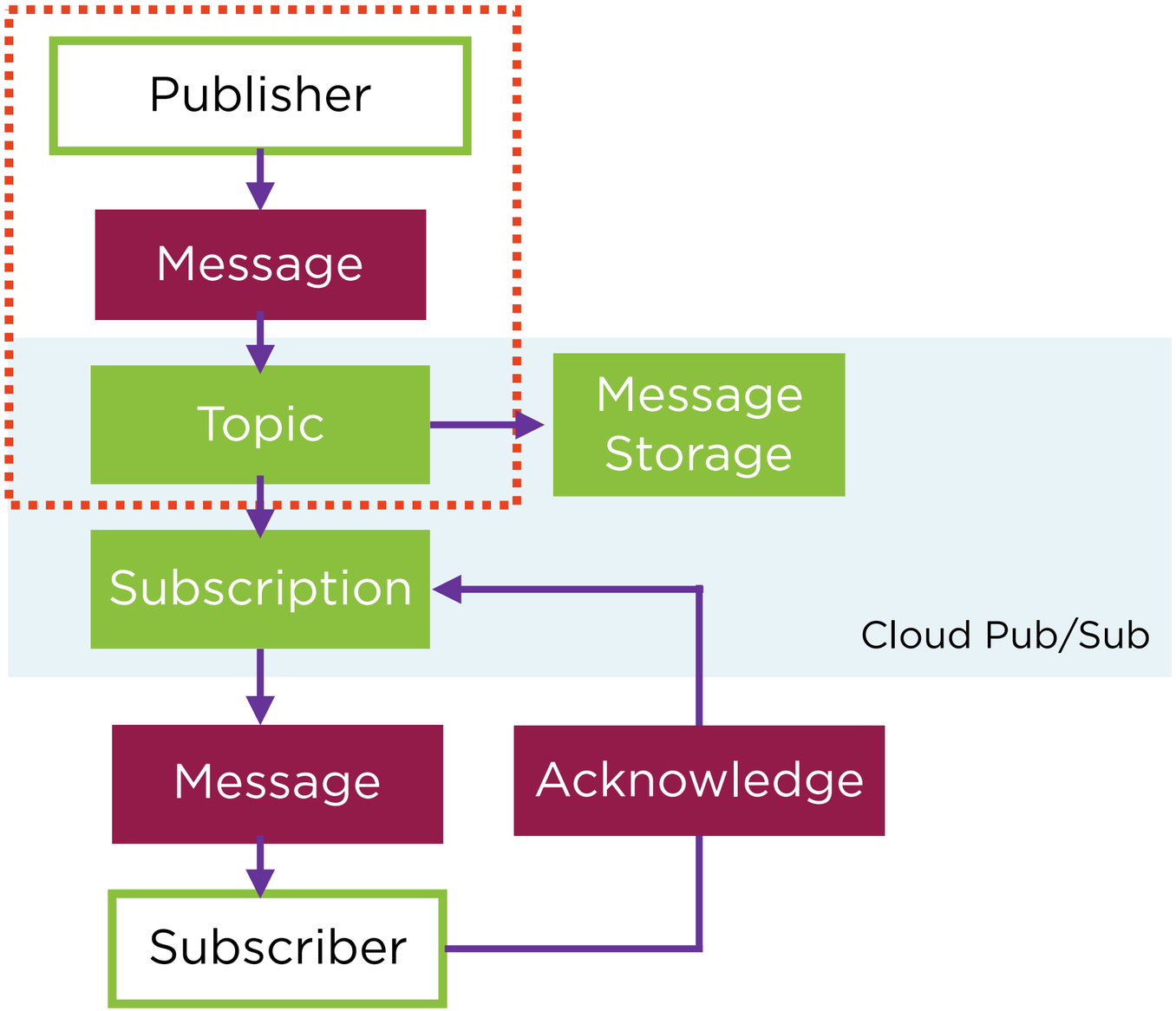
Pub/Sub

Messaging “middleware”

Many-to-many asynchronous messaging

Decouple sender and receiver

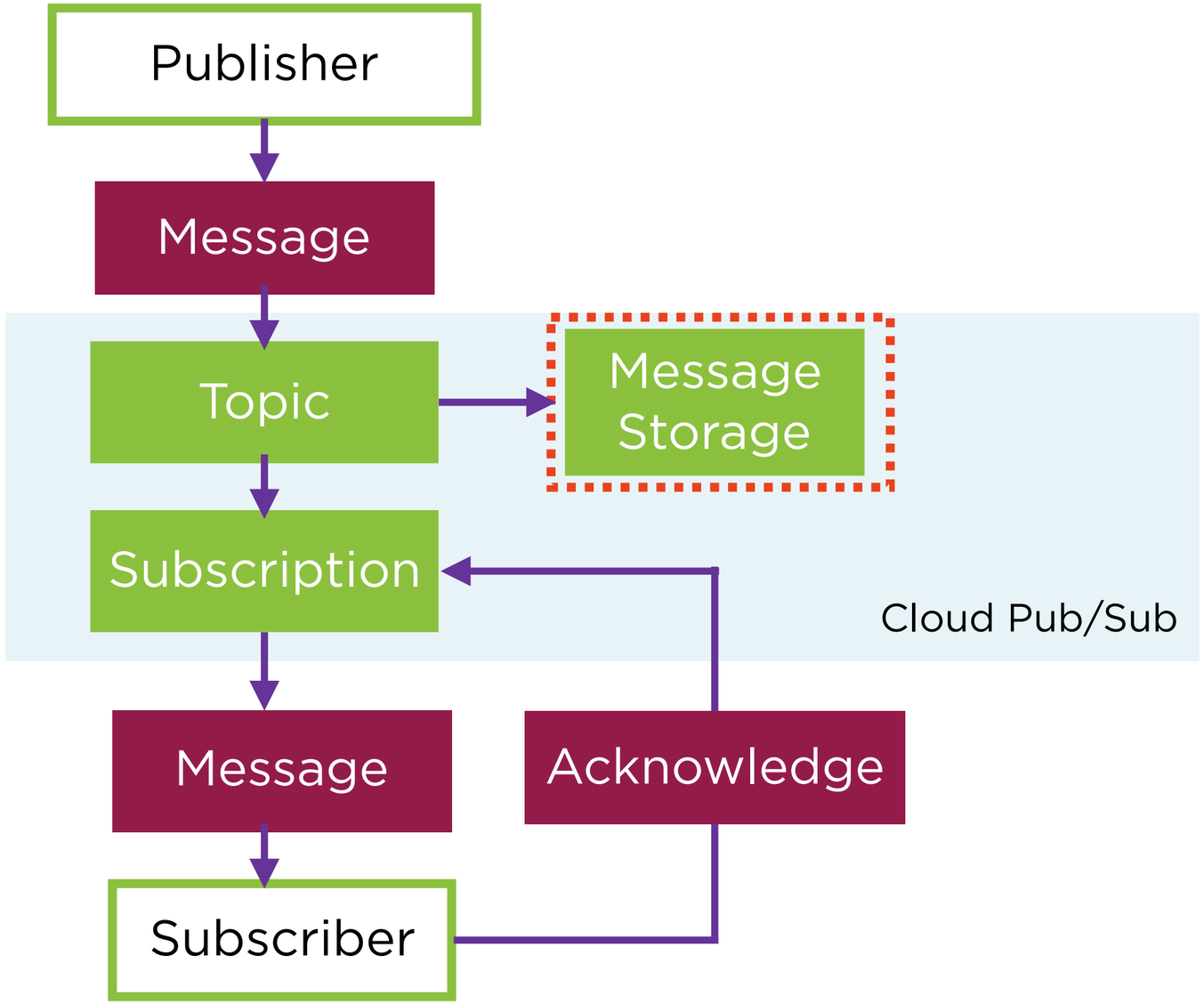
Pub/Sub Messaging



Publisher creates topic and sends messages to the topic

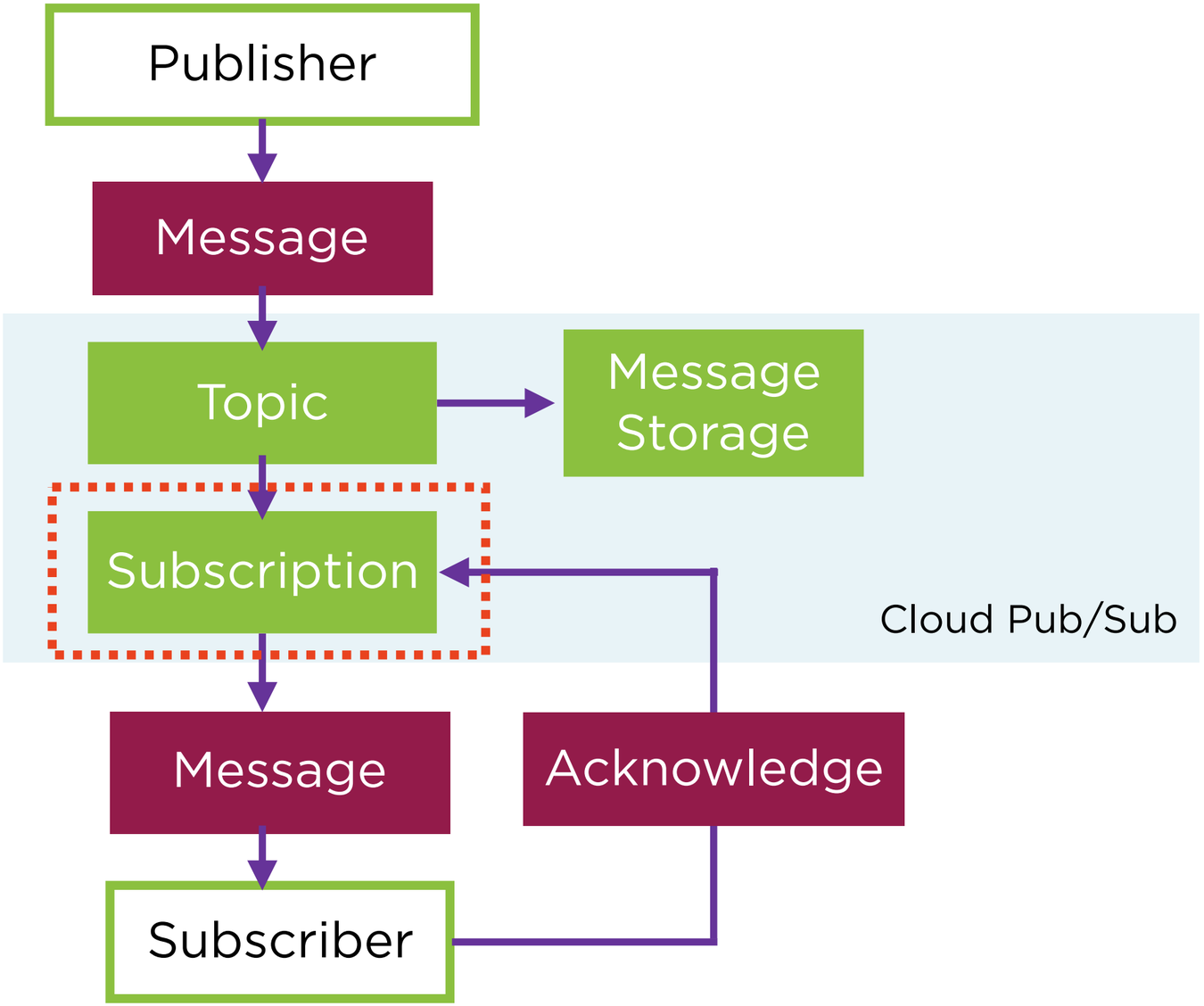
Message contains payload and optional metadata attributes

Pub/Sub Messaging



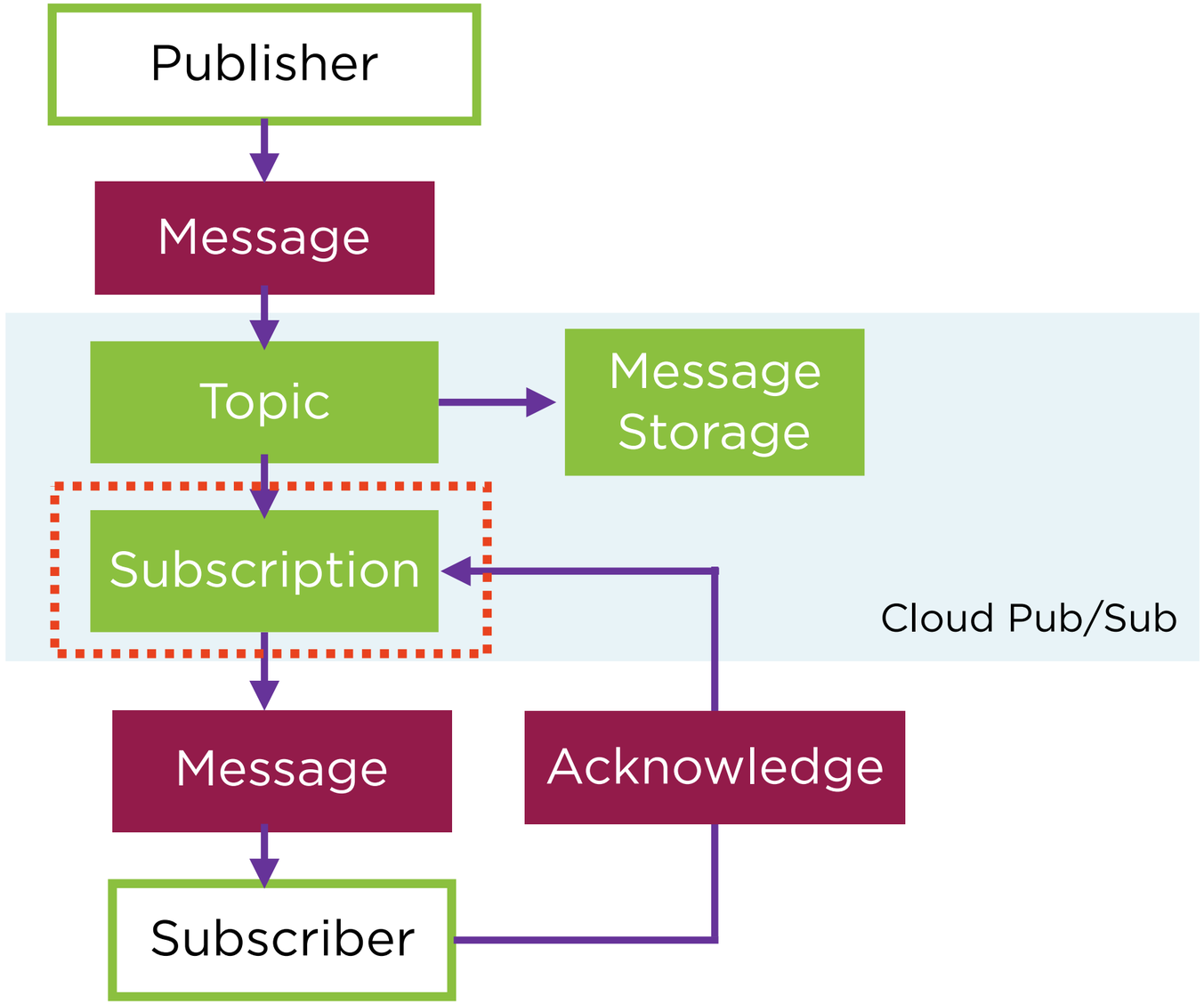
Messages stored till they are delivered to subscribers

Pub/Sub Messaging



Pub/Sub forwards messages from a topic to subscribers

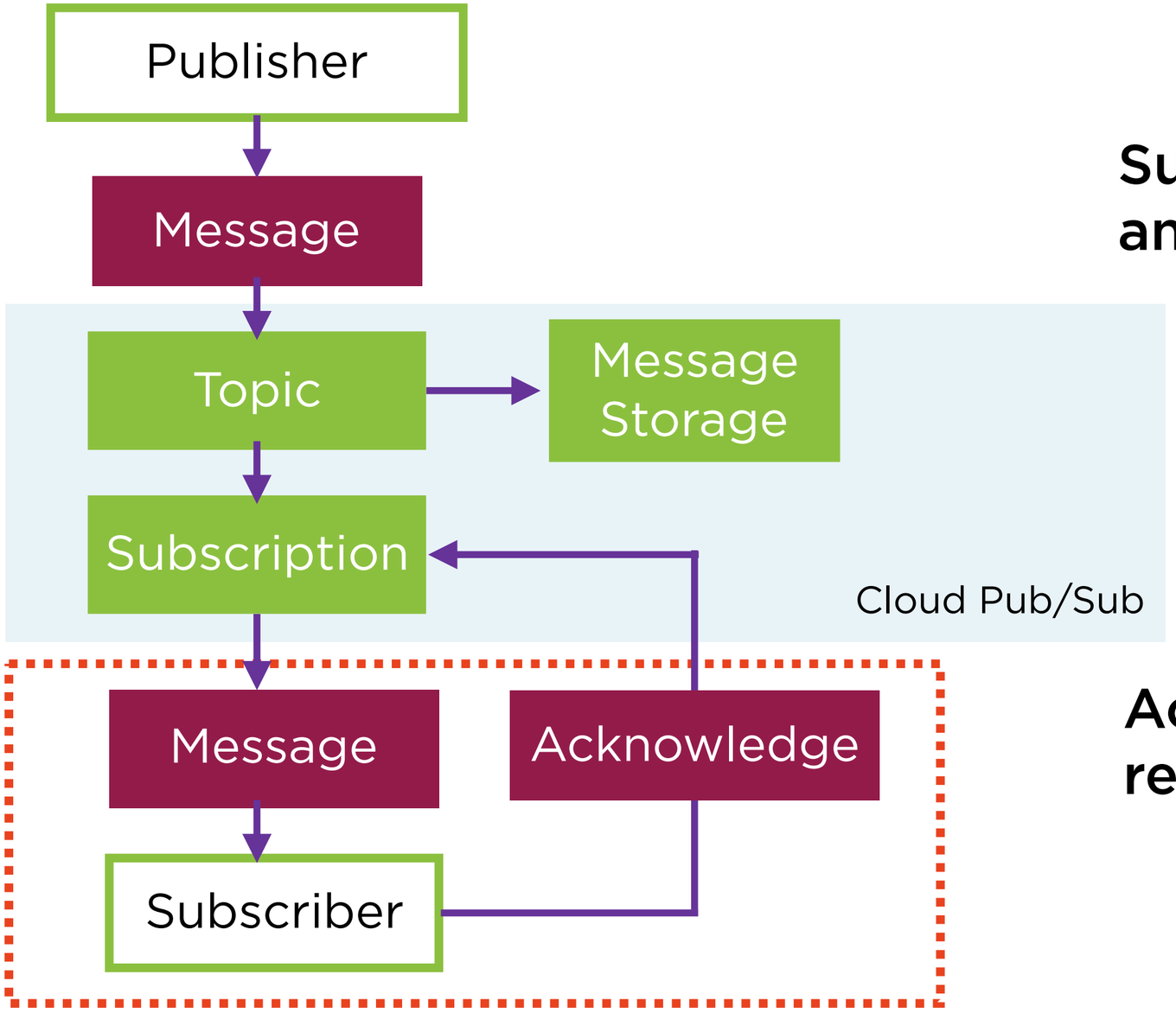
Pub/Sub Messaging



Subscribers pull messages from Pub/Sub

Or are pushed messages from Pub/Sub to the subscriber's endpoint

Pub/Sub Messaging



Subscribers receive messages and acknowledge them

Acknowledged messages are removed from the Pub/Sub queue

Demo

Defining pipelines using templates

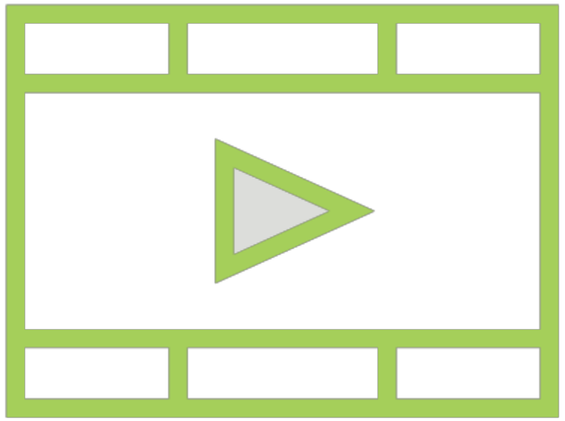
Summary

Templates for ease of pipeline creation

Deploying pipelines from templates

Working with streaming data using Pub/Sub

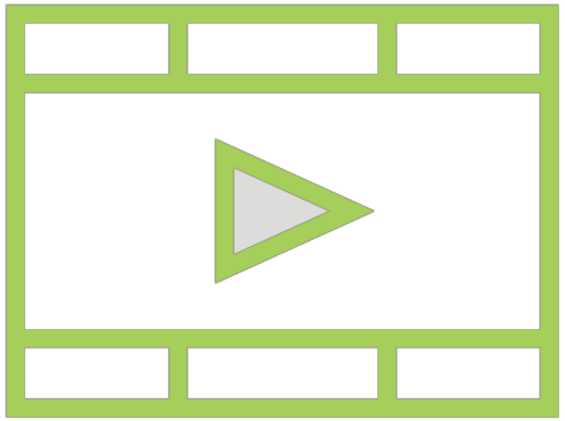
Big Data Processing



The Building Blocks of Hadoop - HDFS, MapReduce, and YARN

Getting Started with Spark 2

Big Data Processing on Cloud Platforms



Architecting Big Data Solutions Using Google Dataproc

Big Data on Amazon Web Services