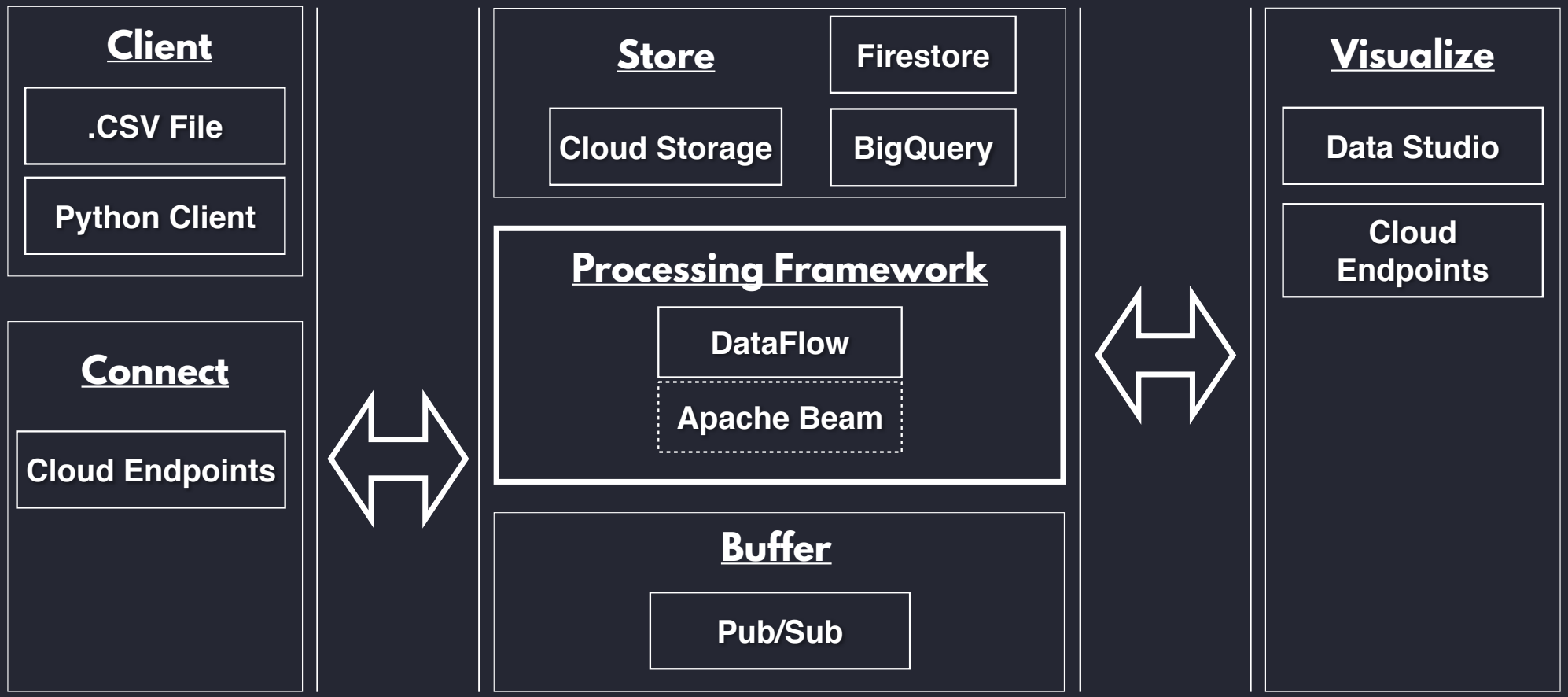
Platform Design

Data Engineering on GCP



Client

- Python Client
  - Reads .csv

data.csv

```
InvoiceNo,StockCode,Description,Quantity,InvoiceDate,UnitPrice,CustomerID,Country
536365,85123A,WHITE HANGING HEART T-LIGHT HOLDER,6,12/1/2010 8:26,2.55,17850,United Kingdom
536365,71053,WHITE METAL LANTERN,6,12/1/2010 8:26,3.39,17850,United Kingdom
536365,84406B,CREAM CUPID HEARTS COAT HANGER,8,12/1/2010 8:26,2.75,17850,United Kingdom
536365,84029G,KNITTED UNION FLAG HOT WATER BOTTLE,6,12/1/2010 8:26,3.39,17850,United Kingdom
536365,84029E,RED WOOLLY HOTTIE WHITE HEART.,6,12/1/2010 8:26,3.39,17850,United Kingdom
536365,22752,SET 7 BABUSHKA NESTING BOXES,2,12/1/2010 8:26,7.65,17850,United Kingdom
```

  - Select data e.g. from today or number of lines

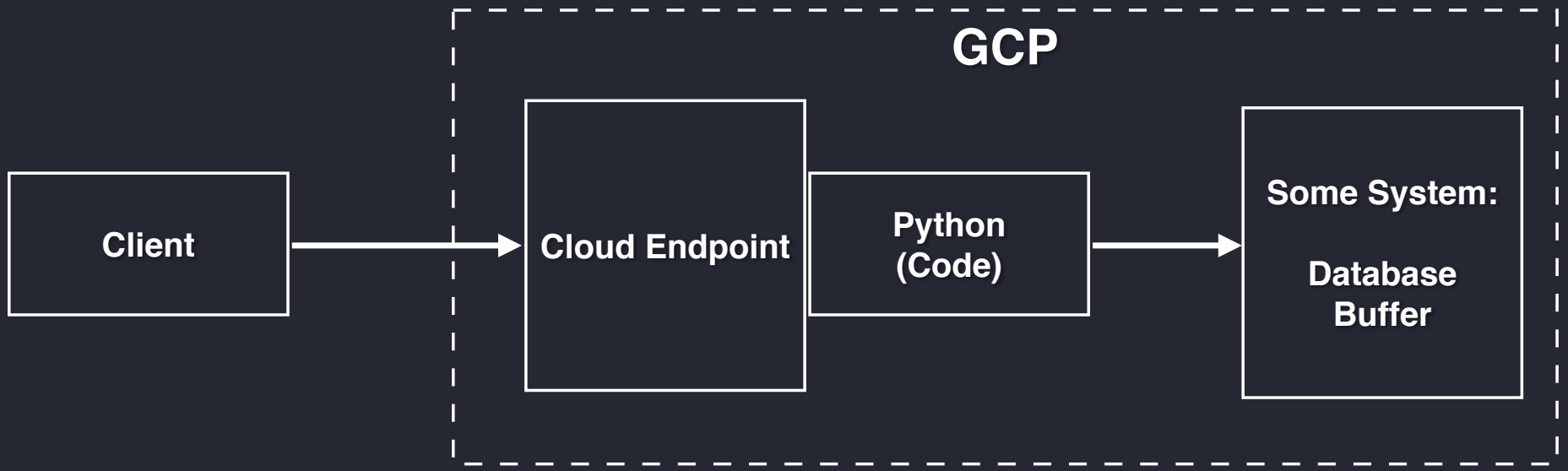
- Transforms each line into JSON string

```json
{
    "InvoiceNo": 536365,
    "StockCode": "85123A",
    "Description": "WHITE HANGING HEART T-LIGHT HOLDER",
    "Quantity": 6,
    "InvoiceDate": "12/1/2010 8:26",
    "UnitPrice": 2.55,
    "CustomerID": 17850,
    "Country": "United Kingdom"
}
```

- Writes each JSON string into sink (API Gateway)

Connect

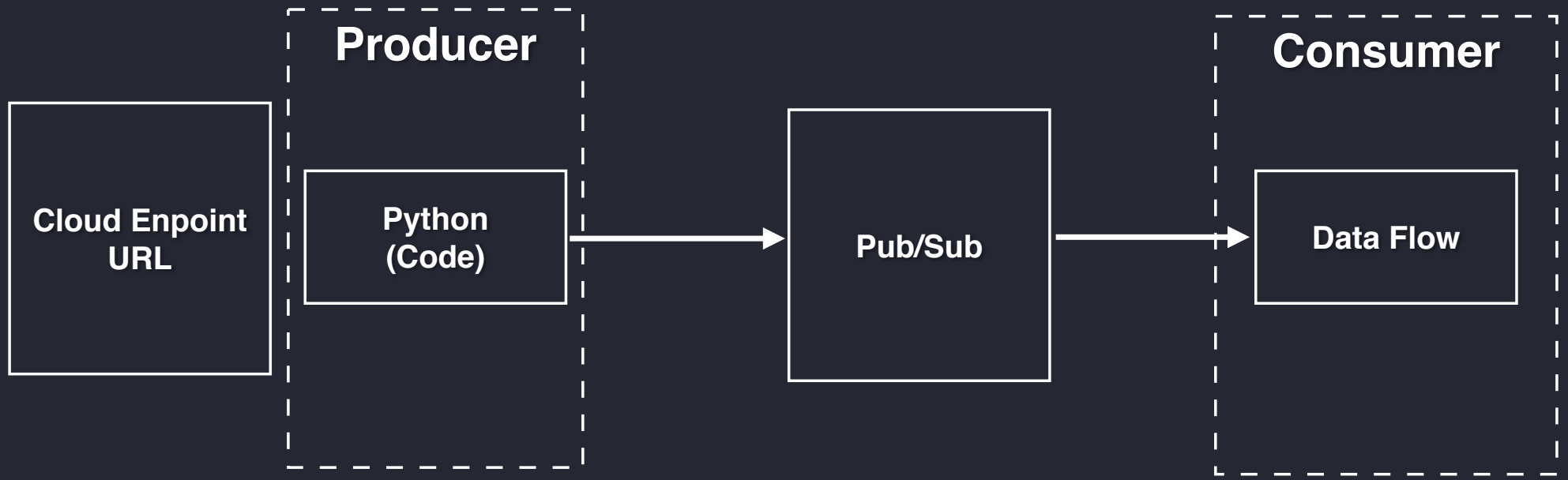
- Cloud Endpoint
- Python



Team Data Science

Buffer

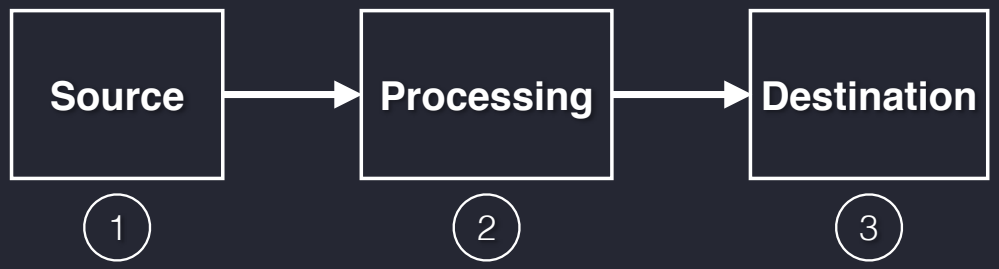
- Pub/Sub

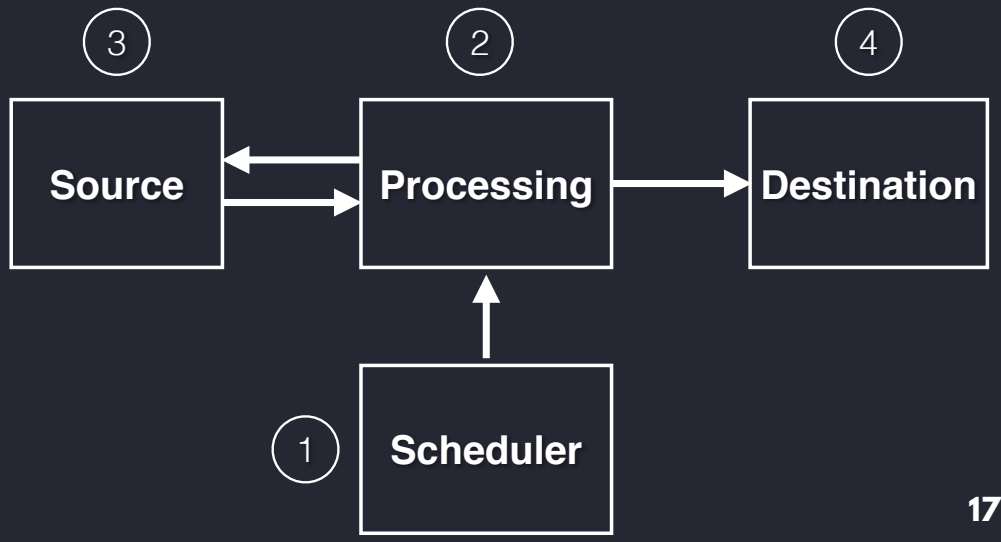


Team Data Science

Process

- Streaming Processing
  - DataFlow with triggers on Source
  - Continuous Process
- Batch Processing
  - DataFlow
  - Cloud Composer for Scheduling





Team Data Science

Store

- Cloud Storage
- Firestore NoSQL
  - Document based Store like MongoDB
  - Transactions
- BigQuery Data Warehouse
  - Analytics Layer
  - Distributed Storage And Processing

Visualize

- APIs
  - Access for Apps, UIs..
  - Execute Queries and Transactions
  - Simple, Stateless
- Tableau
  - Business Intelligence Tool
  - Installed On Your PC
  - Connects to Redshift

Data Pipelines

Data Ingestion Pipeline

- Client
  - Simulates Streaming
  - Sends CSV Rows as JSON
- Cloud Endpoint API
- Python Code For Write Into Pub/Sub
- Pub/Sub buffer for streaming

Team Data Science

Stream to Cloud Storage Pipeline

- Pub/Sub insert triggers Dataflow
- Dataflow waits for some time
  - Writes all messages in queue to Cloud Storage as file

Stream to Firestore Pipeline

- Pub/Sub Insert triggers Dataflow insert for Firestore
  - Dataflow reformats/preprocesses messages
  - Dataflow writes customer data (customer + invoices)
  - Dataflow writes invoice data (invoice + stockcode)

Team Data Science

Visualization Pipeline API

- APIs for UI (Items in Invoice)
  - Data rests in Firestore table Invoices
  - Client requests Items for InvoiceNo (Request parameter)
  - Python of Cloud Endpoint triggered by API queries Firestore with InvoiceNo

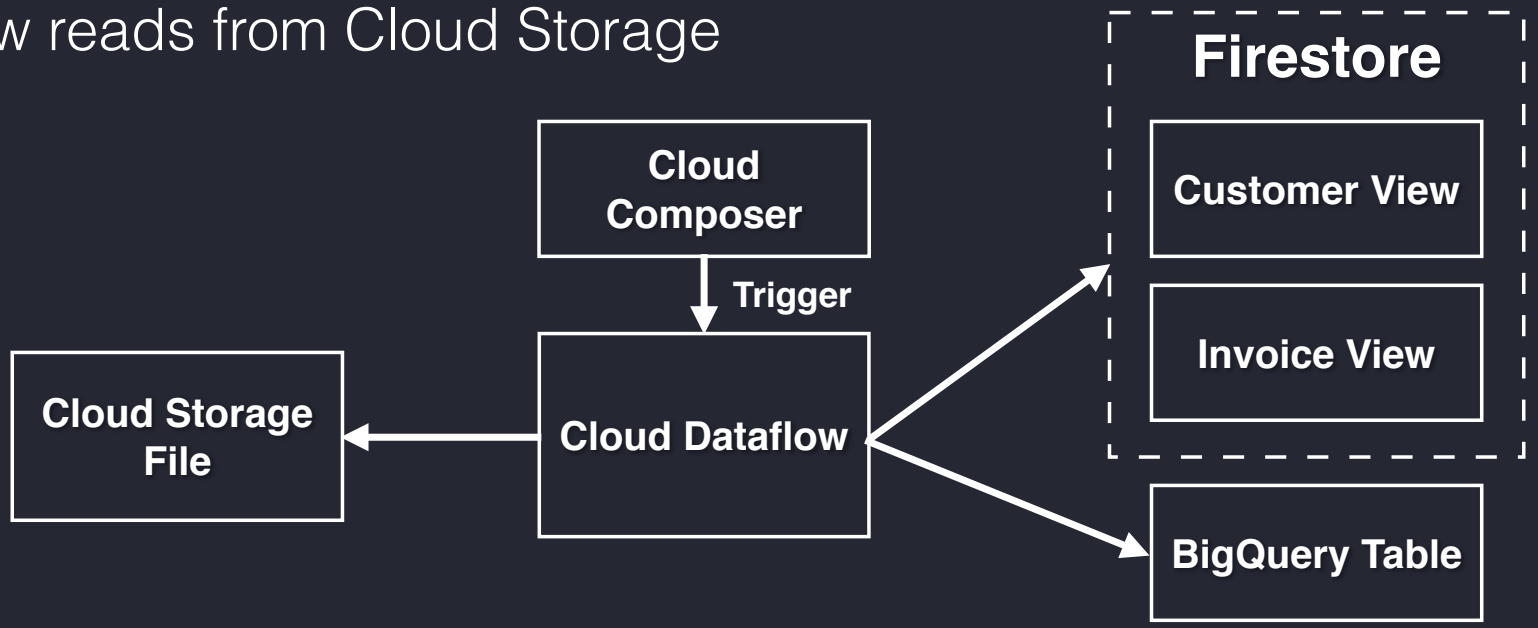
Visualization Pipeline BigQuery Data Warehouse

- Pub/Sub triggers Data Flow
- Data Flow writes messages into BigQuery

Team Data Science

Batch Processing Pipeline

- Bulk import Pipeline
- Triggered through Cloud Composer
- Dataflow reads from Cloud Storage
- Writes data into Firestore
- Writes into BigQuery



Team Data Science