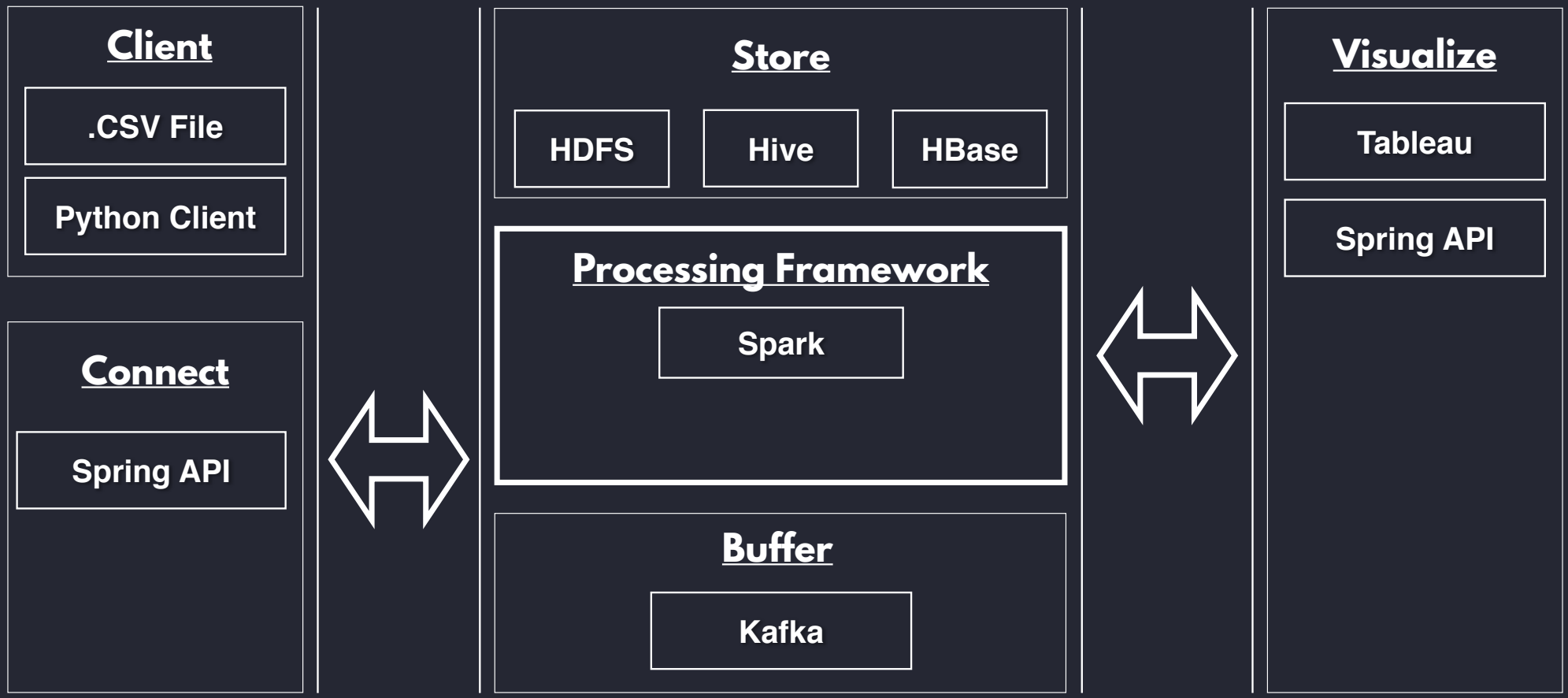
Platform Design

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

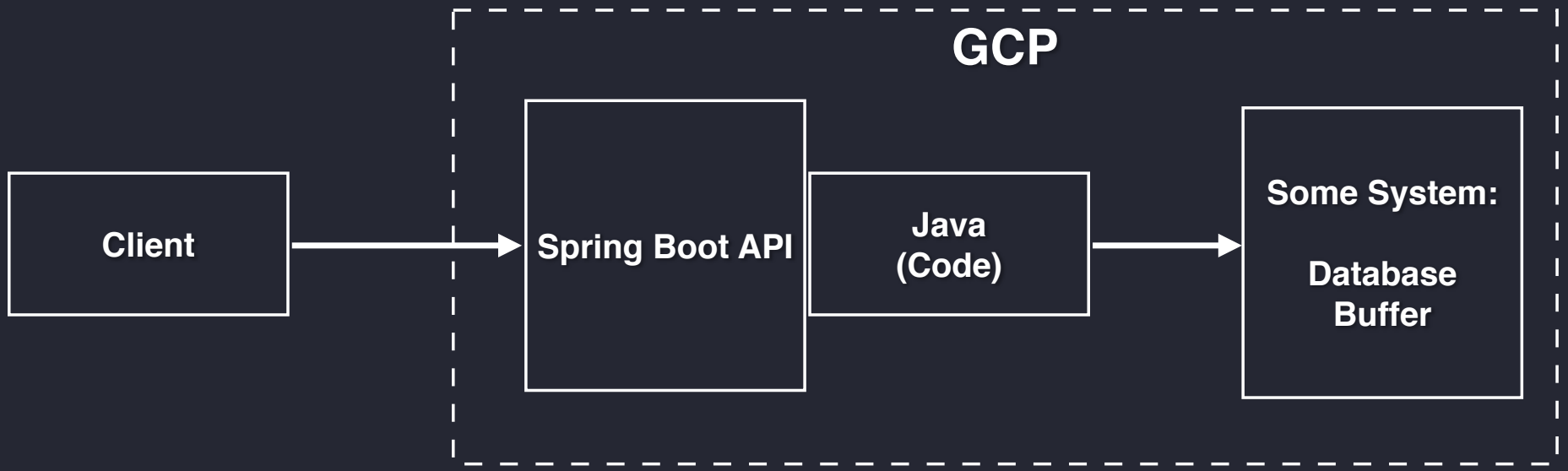
- Transforms each line into JSON string

```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

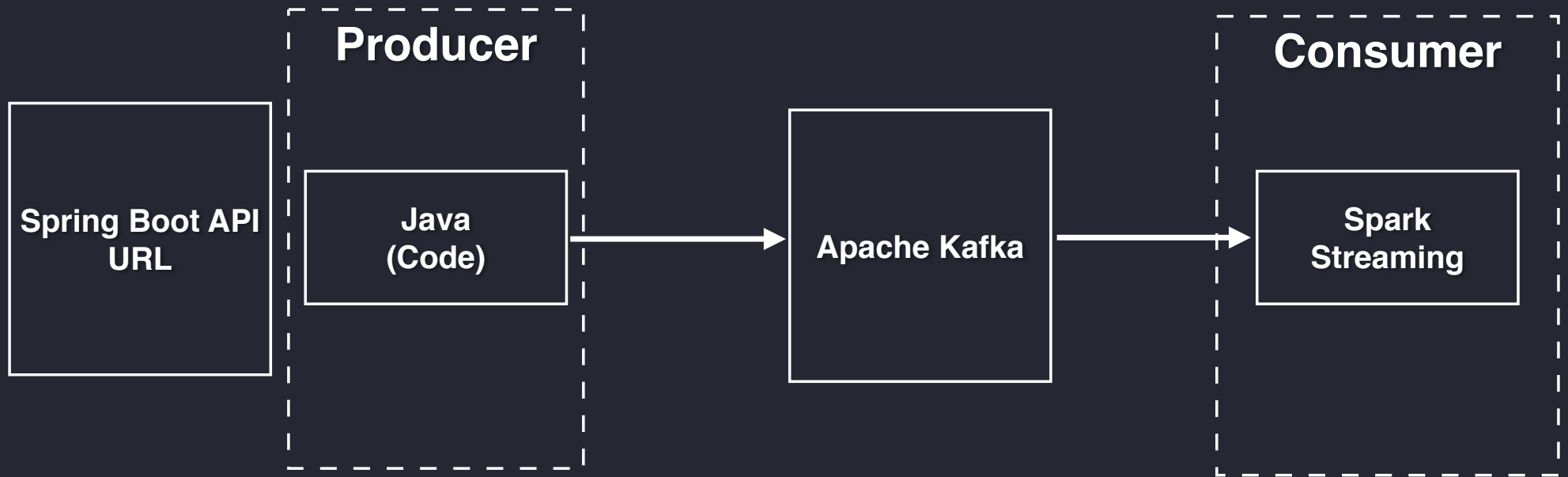
- Spring Boot
- Java



Team Data Science

Buffer

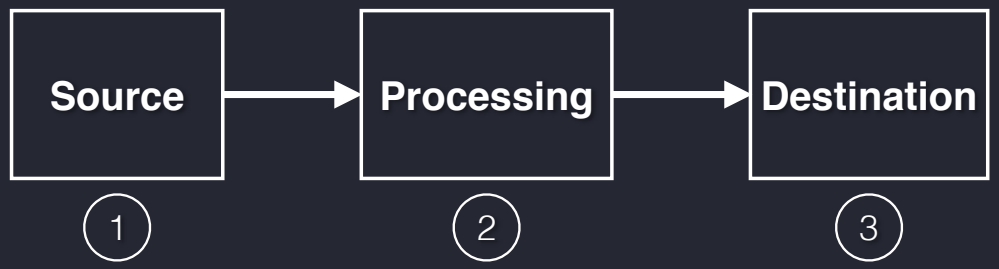
- Apache Kafka



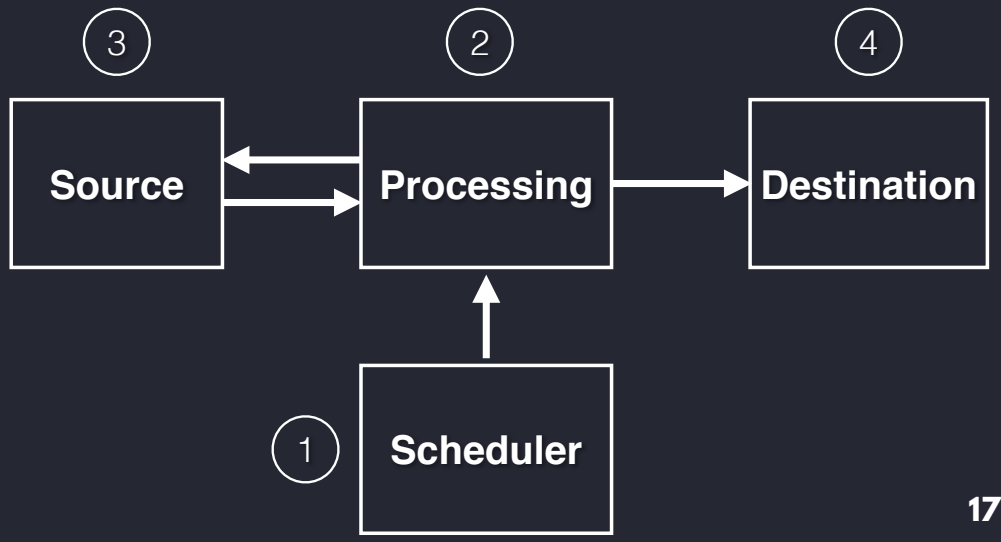
Team Data Science

Process

- Streaming Processing
  - Spark Streaming with triggers on Source
  - Continuous Process



- Batch Processing
  - Spark Batch Processing
  - Oozie for Scheduling



Team Data Science

Store

- HDFS (Hadoop File System)
- HBase NoSQL
  - Key Value Store
  - Transactions (possible but manual coding needed)
- Hive Data Warehouse
  - Analytics Layer processing with Spark
  - Can access data from HDFS or HBase

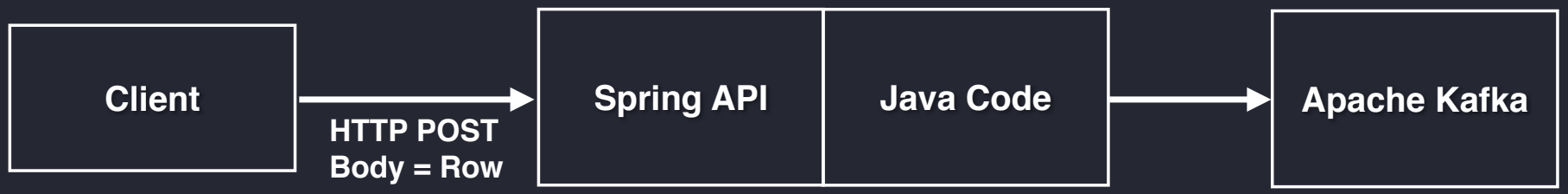
Visualize

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

Data Pipelines

Data Ingestion Pipeline

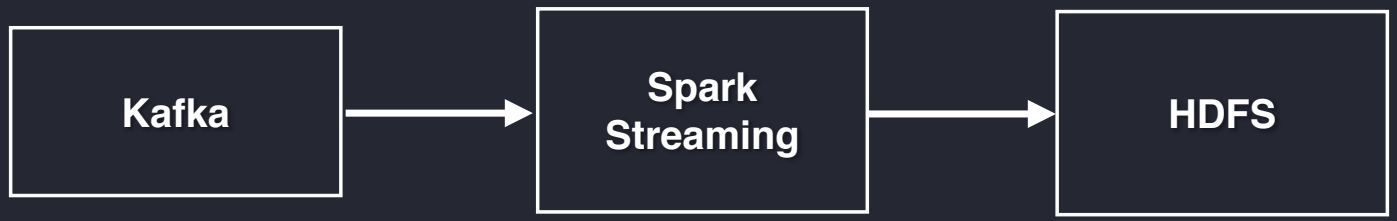
- Client
  - Simulates Streaming
  - Sends CSV Rows as JSON
- Spring API
  - Java Code For Write Into Kafka
  - Kafka buffer for streaming



Team Data Science

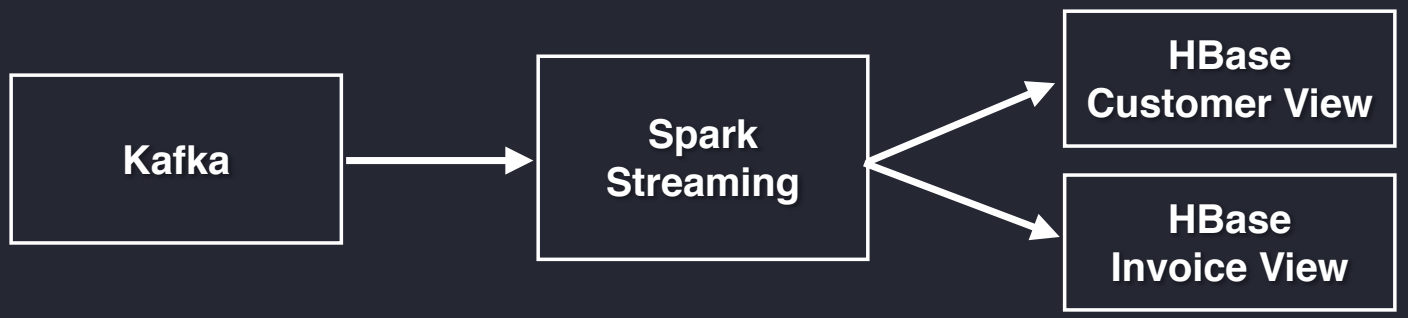
Stream to HDFS Pipeline

- Kafka insert gets fetched by Spark Streaming job
- Spark Streaming micro batches run in small intervals
  - Microbatch writes all messages in Kafka to HDFS as file



Stream to HBase Pipeline

- Kafka messages get processed by Spark
  - Spark reformats/preprocesses messages
  - Spark writes customer data (customer + invoices)
  - Spark writes invoice data (invoice + stockcode)



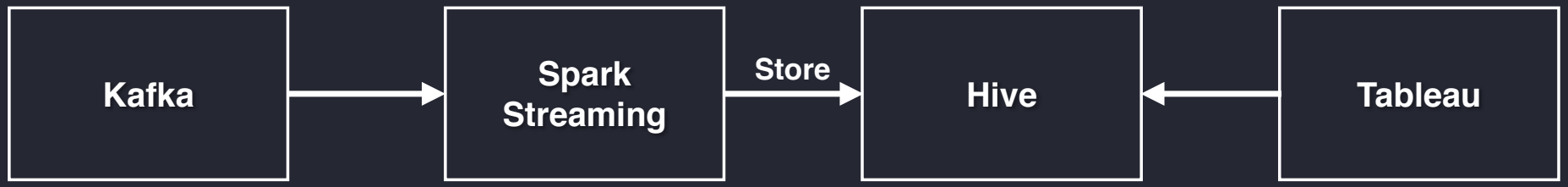
Team Data Science

Visualization Pipeline API

- APIs for UI (Items in Invoice)
  - Data rests in Hbase table Invoices
  - Client requests Items for InvoiceNo (Request parameter)
  - Java code of Spring API queries HBase with InvoiceNo

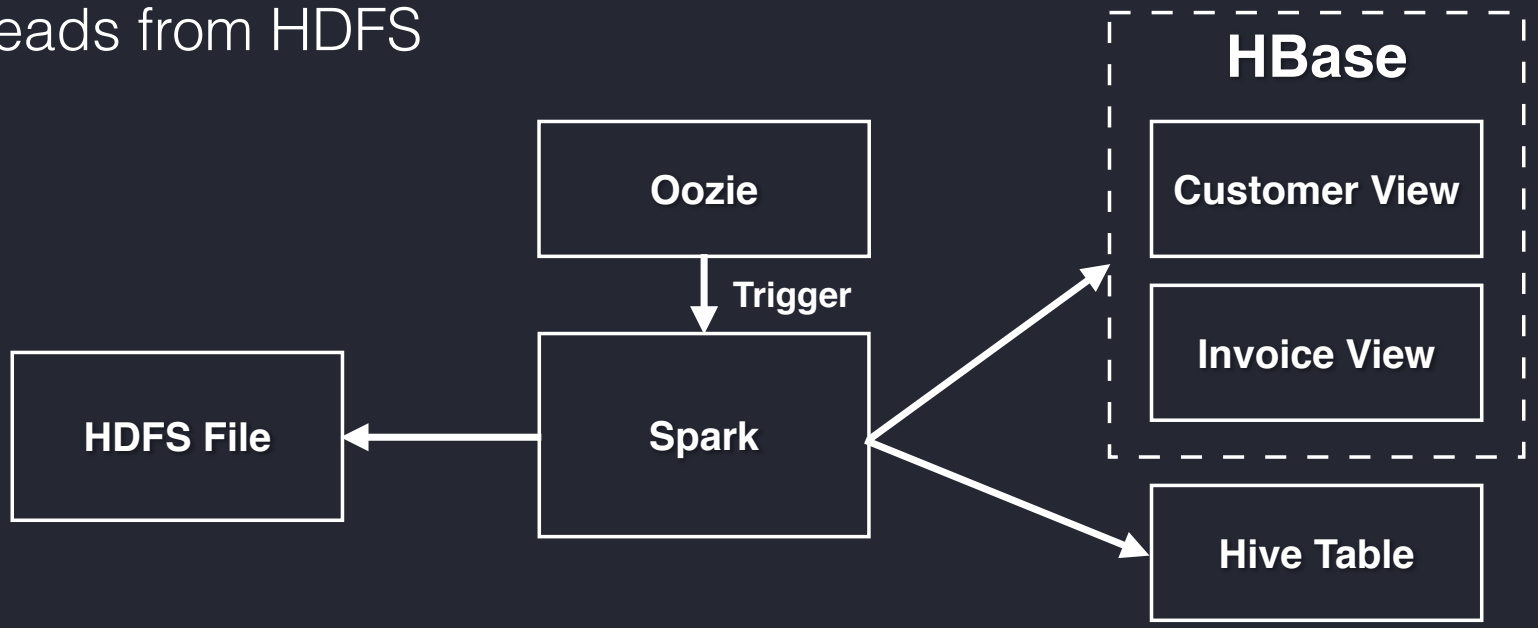
Visualization Pipeline Hive Data Warehouse

- Messages are in Kafka
- Spark Streaming writes messages into Hive
- Visualize with Tableau



Batch Processing Pipeline

- Bulk import Pipeline
- Triggered through Oozie
- Spark reads from HDFS
- Writes data into HBase
- Writes into Hive



Team Data Science