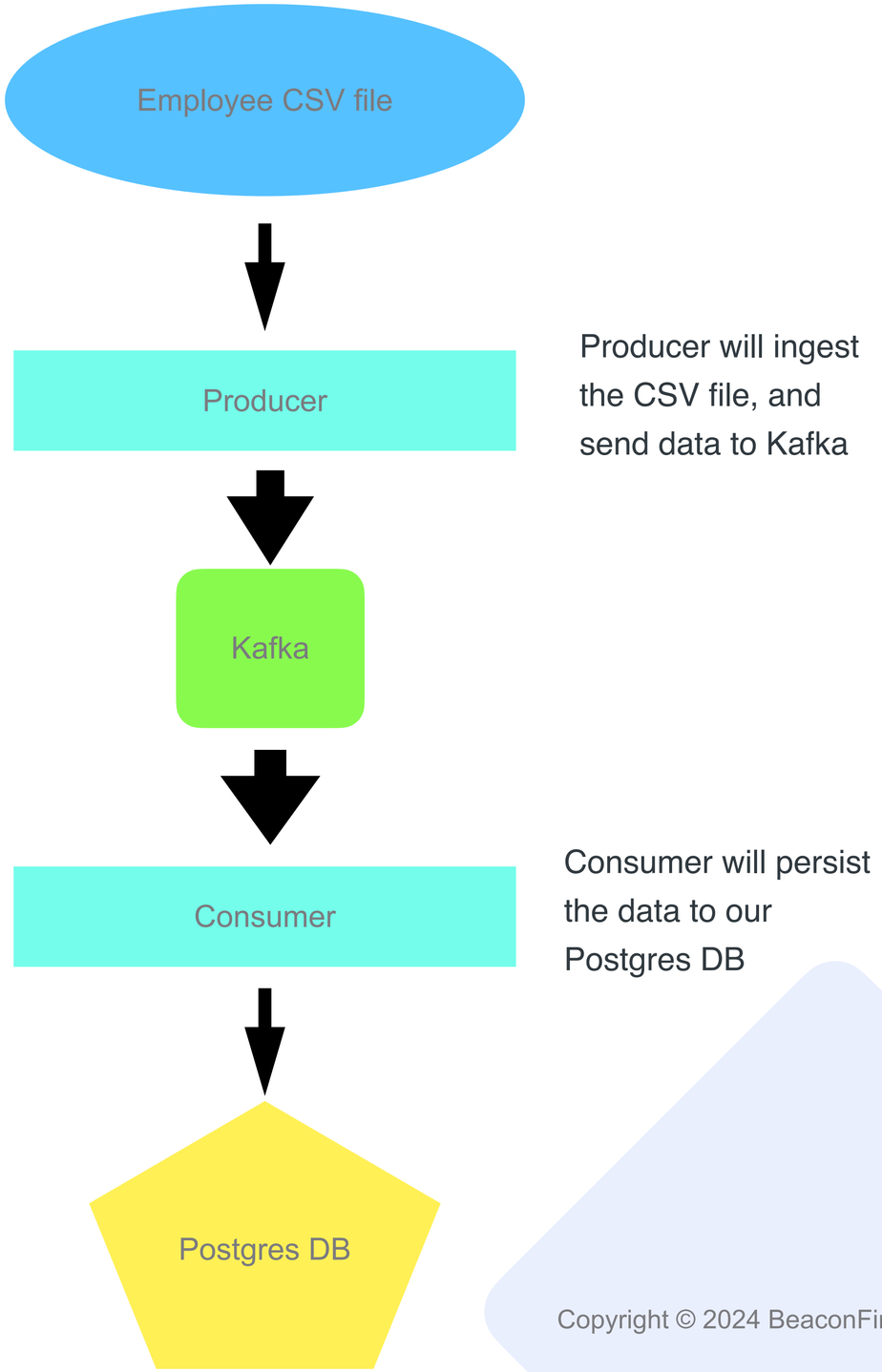
Kafka Project 1

Kafka Project 1: ETL pipeline

- This project will emphasize on creating a data pipeline in which CSV files are being generated on a regular basis, and then data is being cleaned, transformed and then later persisted into some storage for data analytics.



Data Producer

DataProducer Responsibilities - (Extract + Transform)

- Ingest Employee_Salaries.csv file (will be in resources folder)
- Perform these transformations -
  - Ingest only these Departments -
    - ECC
    - CIT
    - EMS
  - Round off the Salary to lower number
  - Employees hired after 2010
- Send this data to Kafka

Data Consumer

DataProducer Responsibilities - (Load)

- Ingest the data into the **Department_Employee** Table which will have this schema -
  - department_division: varchar
  - position_title: varchar
  - hire_date: Date
  - salary: int32
- With every message, also update the total salary given by each department. Schema for the department table -
  - department: varchar
  - total_salary: int64

Table Schema

```
CREATE TABLE department_employee(

department VARCHAR(100),

department_division   VARCHAR(50),

position_title VARCHAR(50),

hire_date DATE,

salary decimal

);
```

```
CREATE TABLE
public.department_employee_salary (

department varchar(100) NOT NULL,

total_salary int4 NULL,

CONSTRAINT department_employee_salary_pk PRIMARY KEY (department)

);
```

helper code

```
#update statement to update the table

cur2.execute(
    "insert into  department_employee_salary  (department,total_salary)
  values (%s,%s) on conflict(department) do update set total_salary =
  department_employee_salary.total_salary + %s    ",
   (e.department, int(float(e.salary)), int(float(e.salary))))
```

Success Criteria

1, Set up the Kafka docker image
docker-compose build
docker-compose up -d

2, Open two new terminals, navigate to your project folder in both terminals. In one terminal run:
python producer.py

And in the other, run:
python consumer.py

Depending on your set up, your producer and consumer may keep running/stop automatically. Use CTRL+C to force kill the process.

3, You should get the same results.

department_employee_salary | Enter a SQL expressio

	department	total_salary
1	EMS	3,779,570
2	CIT	9,102,142
3	ECC	2,042,698