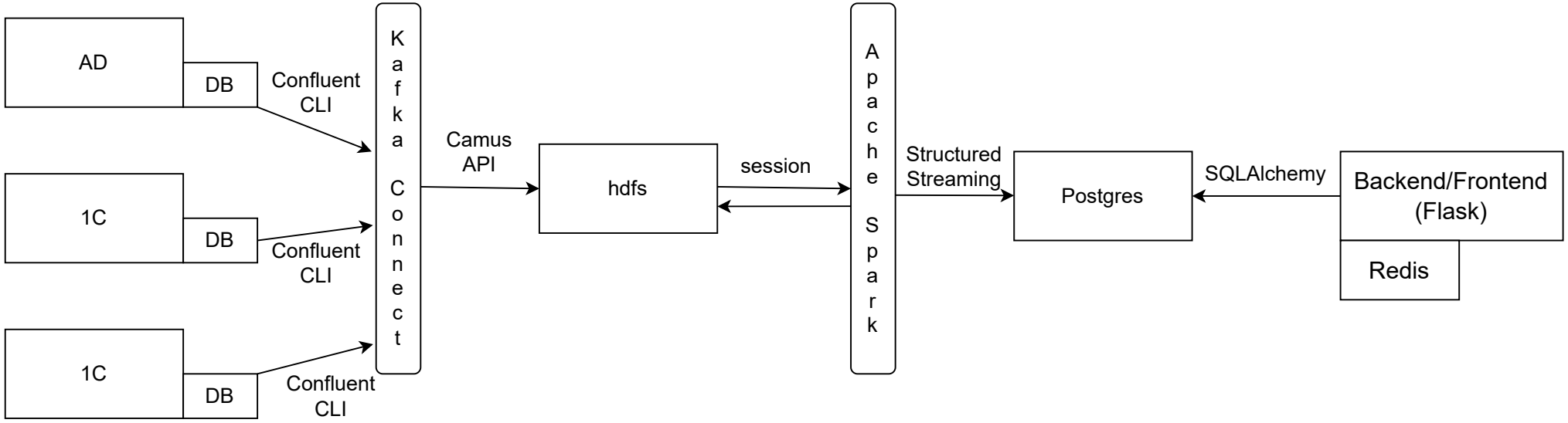

Case 1

Пусть есть 3 системы, собирающие информацию о сотруднике в разном формате.
Данные собираются с помощью kafka connect'a с confluent cli для настройки db. Далее с помощью camus api (MapReduce job'a) переливаем сырые данные в hdfs. Благодаря спарку преобразуем данные из hdfs и сохраняем их в postgres. Далее система для работы с данными это поднятые сервисы (вместе с поднятым redis'ом если требуется), которые запрашивают данные напрямую из бд. Считаю это разумным т.к. обновление происходит только при найме/увольнении и кол-во данных не меняется слишком).

Холодное хранилище - hdfs.
Теплое - postgres.
Горячее - redis как аналог кеша если требуется.

Репозиторием является postgres с предобработанными и приведенными к единой структуре данными благодаря spark.