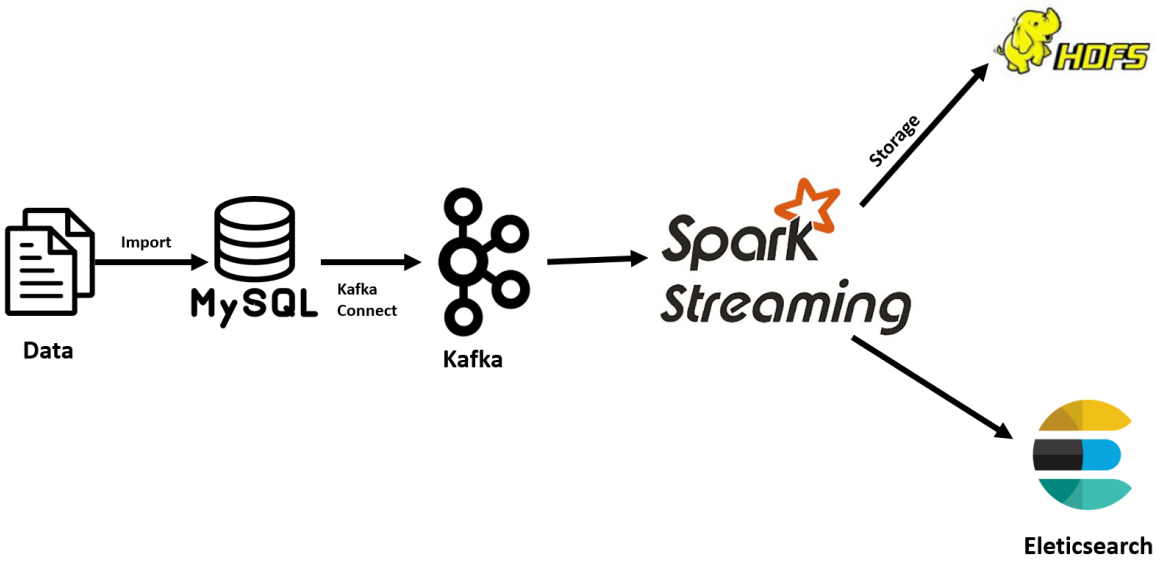
1. Architecture



2.Data

Data là dữ liệu thông tin các chuyến bay được ghi lại theo từng năm ở Mỹ.Gồm hơn 1 triệu bản ghi.
(source : Data Expo 2009: Airline on time data - ASA Statistical Computing Dataverse (harvard.edu))

Data gồm các trường:

- Year:năm của chuyến bay.
- Month:tháng của chuyến bay
- DayOfMonth:ngày trong tháng (1 đến 31)
- DayOfWeek:ngày trong tuần
- DepTime:thời gian khởi hành thực tế
- CRSDepTime:thời gian khởi hành theo lịch trình
- ArrTime:thời gian đến thực tế
- CRSArrTime:thời gian đến theo lịch trình
- UniqueCarrier: ID nhà cung cấp dịch vu
- FlightNum: số chuyến bay

- TailNum:số đuôi của máy bay
- ActualElapsedTime:thời gian thực tế đã trôi qua của chuyến bay, tính bằng phút
- CRSElapsedTime: thời gian trôi qua theo lịch trình của chuyến bay, tính bằng phút
- AirTime;thời gian trên không của chuyến bay, tính bằng phút
- ArrDelay:đến trễ, tính bằng phút
- DepDelay:khởi hành trễ, tính bằng phút
- Origin:sân bay xuất phát
- Dest:sân bay đến
- Distance:khoảng cách bay.
- Cancelled:trạng thái hủy
- WeatherDelay:hoãn bay do thời tiết, tính bằng phút

Dữ liệu có nhiều trường không cần thiết cho mục tiêu được nêu bên dưới trong quá trình làm em có thể lược bỏ 1 sô trường không cần thiết.

3.Mục tiêu
Sau khi xử lý dữ liệu sử dụng eleticsearch để tra các thông tin cần thiết với tốc độ cao.

Sử dụng Apache Spark và Java MapReduce để đưa ra số liệu sân bay cho hạ cánh nhiều nhất,tỉ lệ hoãn hủy chuyến ở mùa nào là nhiều nhất.