

런드리고 스노우플레이크로 실시간 데이터 수집·분석 세탁물 수거 요청 예측정확도 향상

LaundryGo

국내 대표적인 세탁 서비스 런드리고는 가장 세탁 요청이 많은 강남구의 수요 지표를 확인하기 위해서 실시간 접속 데이터를 활용, 강남구의 물량을 예측하고 있다. 이 과정에서 보다 정확한 지표를 쉽게 뽑아내기 위해 스노우플레이크를 도입했다.

런드리고는 2019년 론칭한 국내 대표적인 세탁 서비스다. 서울과 경기 지역에서 직접 세탁을 하는 공장 ‘스마트팩토리’를 세 곳 운영하면서 신청 후 만 24시간 안에 깨끗해진 빨랫감을 돌려주는 일을 한다. 세탁 외에 수선, 고가의 의류를 케어하는 프리미엄 서비스를 제공한다. 월정액 구독이라는 비즈니스 모델을 갖고 있으며, 현재 총 56만 가구의 회원을 확보했다. 그간 220만 건의 주문과 1400만 건의 세탁 경험을 쌓았다. 시리즈C까지 투자받았으며, 누적 유치액은 1225억 원이다.

런드리고의 데이터 분석 역사

세탁은 모두에게 필수이므로, 런드리고 외에도 이 시장에 뛰어들어 경쟁 기업은 더 있다. 런드리고가 가진 차별점은 ‘배송’이다. 미리 주문한 세탁박스(런드렛)에 세탁물을 넣고 스마트폰 앱을 통해 세탁을 신청하면 당일 밤에 수거, 다음날 새벽까지 세탁이 끝난 의류를 돌려받을 수 있다. 만 24시간 안에 깨끗한 세탁물을 받을 수 있게 하는 배송의 흐름이 런드리고가 가진 가장 큰 경쟁력이다. 그러나 이런 배송 운영은 갑작스레 늘렸다 줄였다 할 수 있는 것이 아니므로 데이터를 통한 수요 예측이 매우 중요하다.

런드리고는 고객이 어떤 품목을 맡기고 있는지, 실제 어느 지역의 사람들이 세탁을 많이 맡겨 회사의 매출에 도움이 되는지 등의 지표를 보고 싶었다. 그러나 스타트업의 속성 상 사업 초기에는 서비스 영위에 모든 역량이 집중되므로 데이터 분석은 그 중요도가 뒤로 밀리는 경우가 잦다. 런드리고 역시 마찬가지로 상황이었었는데, 따라서 초반에는 오픈소스 비즈니스 인텔리전스(BI) 도구인 메타베이스만 도입, 볼 수 있는 가능한 지표를 임시적으로라도 확인하는 수준에 머물렀다.

이 기간, 런드리고는 앱을 통해 발생한 여러 데이터를 아파치 카프카나 아마존 RDS, 아마존 S3 등에 적재, SQL 작성을 통해 지표를 확인했다. 그러나 RDS 데이터를 기반으로 지표를 뽑는다는 것이 쉬운 일이 아니었다. 서비스가 계속 업데이트 되면서 변경·추가되는 데이터의 양이 상당한데 그런 변화를 다 쫓아가면서 지표를 산출하기 위한 SQL 작성 작업이 만만치 않았기 때문이다. 대부분 개발자의 도움을 받아야만 SQL을 작성할 수 있어 지표를 보는 데 비효율적이라는 지적이 내부에서 나왔다.

이것이 런드리고가 스노우플레이크를 도입한 이유다. 런드리고는

“일반적인 BI나 RDS 운영만으로는 비즈니스를 고도화하기 위한 지표를 뽑기 어렵다. 그러나 스노우플레이크 도입 후 ‘데이터 드라이브 원스톱 쇼핑(Data Driven one-stop shopping)’이라고 표현할 수 있을 만큼 스마트팩토리에서 일어난 여러 지표를 쉽게 뽑을 수 있게 됐다. - 장성재 런드리고 최고기술책임자(CTO)

스노우플레이크 도입 후, 실시간 데이터를 입수하고 필요한 지표를 추출하고 있다.

스노우플레이크를 도입한 후에는 ▲스노우파이프 스트리밍 ▲스노우파이프 ▲변경 데이터 캡처(CDC) 등의 방식으로 실시간 데이터를 수집하고 있다. 스노우파이프 스트리밍은 실시간 데이터 수집 기능으로 API가 호출되면 적재되는 로그나, API 호출을 통해 나오는 사용자의 액션 같은 것을 카프카에 다 집어 넣는 것이다. 스노우파이프는 RDS로부터 데이터를 읽어 에어플로우를 이용해 데이터를 입수하는 방식이다. 장성재 런드리고 최고기술책임자(CTO)는 “스노우파이프는 S3에 있는 데이터를 내가 원하는 시점에 넣을 수 있다”면서 “S3에 지표 산출이 필요한 기본 데이터를 저장할 때 그 저장 주기를 굉장히 짧게 지정할 수 있으며, 스노우파이프를 짧은 시간 간격으로 실행시키면 거의 실시간에 가까운 데이터 입수가 가능하다”고 말했다.

또 기존 방식으로 캡처되거나 형식이 지정되지 않는 반정형 데이터의 경우 제이슨(JSON)을 많이 활용하는 상황이었다. 런드리고 시스템에서는 로그 형태가 제이슨으로 되어 있으므로 데이터를 스노우파이프 스트리밍을 통해 그대로 입수시켰다.

이렇게 입수된 데이터를 가지고 원하는 지표를 봐야 한다. 런드리고가 주로 보는 데이터는 수거 신청이다. 세탁 서비스에서 수거 신청은 커머스 플랫폼의 ‘주문’과 같다. 쿠팡에서 물건을 주문하면 택배 기사가 상품을 가져다주듯 수거요청을 하면 런드리고에서 배송 기사가 세탁을 수거하러 방문하기 때문이다. 수거 한 건이 하나의 상품 판매가 일어난 것과 동일한 효과를 가진 것은 물론이다.

앞서 언급한 것처럼 RDS로는 지표를 뽑기 어렵다. 그래서 데이터 모델링을 새롭게 했다. 예를 들어 세탁에서 중요한 일자, 시간, 세탁팩토리, 사용자, 주소, 세탁 요금제 등 수거 신청과 관련되어 있을 것으로 예상되는 요소들의 디멘션(dimension) 테이블을 만들고, 이 모델링을 기반으로 수거신청과 관련한 지표를 뽑기 위해 팩트 테이블을 다시 설계했다.

특히 일자과 시간이라는 디멘션은 매우 단순해보이지만, 이는 수거 신청을 하는 시점을 뜻하기 때문에 세탁 서비스에서는 매우 중요한

지표가 된다. 날씨와 요일에 따라 수거량의 차이가 생기기 때문이다. 예를 들어 주말과 계절이 바뀔 때 세탁의 요청이 늘어나고 비가 오면 세탁 요청이 줄어든다. 수거 신청에 날씨나 기온이 영향을 미치는지, 어느 시간대 수거요청이 많은지 등을 분석하기 위해서는 일자와 시간 디멘션이 중요한 기초 데이터에 해당한다. 팩트 테이블과 디멘션 테이블을 적절히 합쳐서 원하는 지표를 쉽게 뽑을 수 있게 됐다.

성과

런드리고는 이와 같은 프로젝트를 통해 데이터 기반으로 강남구 운송물량을 예측할 수 있었다. 왜 강남구나면, 세탁서비스 신청이 가장 많은 곳이기 때문이다. 참고로, 런드리고가 파악한 세탁 주문이 가장 많은 지역은 1위가 강남구, 2위가 서초구, 3위가 마포구 순서다.

예측 정확도는 어떨까? 런드리고 측에 따르면 오차는 ±4.5%다. 예측 물량이 중요한 이유는 런드리고가 실제 수거와 배송 업무를 하기 때문이다. 운송을 갑자기 늘리기는 어려우므로 예측의 정확도가 중요하다. 예측 정확도에는 전날의 운송 물량과 전주 동일 요일의 운송 물량, 그리고 당일 18시까지 신청한 물량(런드리고는 밤 10시까지 당일 세탁 수거 요청을 받는다) 등이 중요한 영향을 끼친다. 운송관점에서는 예측치보다 수거 물량이 많을 경우 곤란한 상황이 오기 때문에 정확한 예측을 통해 그보다 여유 있게 운송 차량을 운영하는 것이 낫다. 런드리고 측은 앙상블 모델을 통해 예측했을 때 조금 더 결과가 잘 나오는 걸 볼 수 있었다고도 조언한다.

향후 계획

세탁을 직접 하는 런드리고의 경우 빨래 공장인 스마트팩토리의 애플리케이션이 있다. 현재 서울과 경기 지역에 세 개의 팩토리를 운영하면서 하루 수만 장의 세탁물을 세탁한다. 그 결과 쌓이는 데이터의 양이 굉장히 많다. 런드리고는 이 데이터를 가지고 세탁 공정의 흐름을 효율화하려 한다. 매일 수만 장의 세탁을 하기 위해서는 세탁물 수거에서부터 등록, 세탁과 사용자 별 세탁물 분류와 배송 등 세탁 공정의 원활한 흐름이 굉장히 중요하기 때문이다.

스노우플레이크에 대하여

스노우플레이크는 데이터 클라우드를 통해 데이터를 가치있게 활용할 수 있도록 합니다. 데이터 사일로 없이 통합된 데이터를 활용할 수 있고, 데이터를 안전하게 공유하며, AI/ML 및 다양한 분석 워크로드를 지원합니다. 2023년 7월 30일 기준, Forbes Global 2000 기업 중 639개 기업을 비롯한 다양한 산업 분야의 8,537개사의 고객이 스노우플레이크 데이터 클라우드를 사용하여 비즈니스를 강화하고 있습니다. 자세한 내용은 www.snowflake.com/ko에서 확인할 수 있습니다.