**Data** :

* 3 tháng ( từ vài năm trước )
* 700.000 sản phẩm\*
* 7 triệu dòng
* Json file

\***Sản phẩm** gồm các thuộc tính:

* ID
* 3 level (ví dụ: quần áo -> đầm -> đầm xòe)
* URL
* Người dùng(lượt click,rating,..)
* Thời gian tương tác từ người dùng.
* ...

Một số **vấn đề** bài toán:

* Người dùng lớn -> scale tăng
* Baseline build ???
* Preprocess data chiếm 85% -> 95%
* Log tương tác khi người dùng nhấp vào sản phẩm.
* Tính chính xác cao: dữ liệu trả về nhanh.
* GUI: không bắt buộc một format chuẩn.