ENGINEERING HONORS PROGRAM
UNIVERSITY OF MICHIGAN

**Engineering Honors Capstone Project**
**MDP Industry-Sponsored (Union Pacific)**

UNION PACIFIC

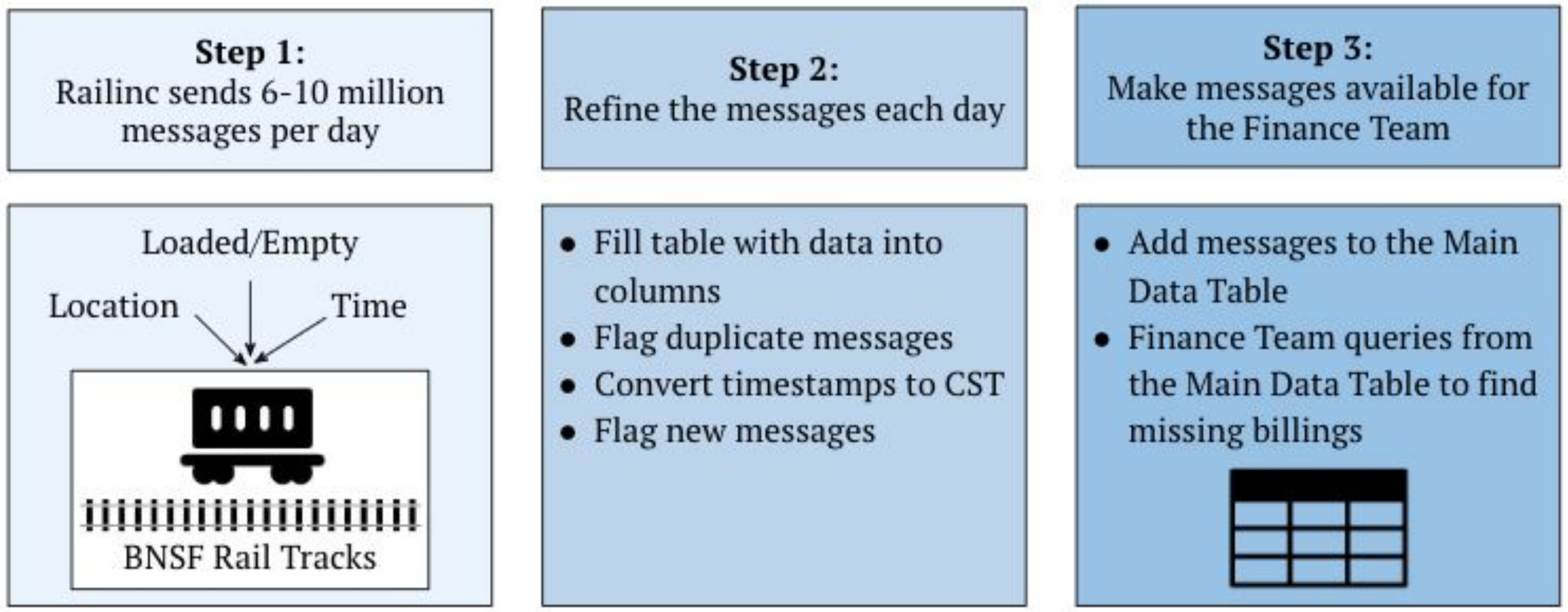
# Big Data Backbone for Advanced Analytics

MDP Team Members: Ann Stone, Celina Pan, Neil Kim, Ken Mahattanadul (Honors Capstone), Conan Wu
Advisor: Prof. William Arthur

## Project Overview:

- *Main Problem*: Union Pacific's Finance Team needs data not available in their Main Data Table
- Finance Team uses railcar data to find missing billings
- Missing data from railcars when on other railroad company's tracks
- Forces manual searching through Railinc** for missing data

** supplier of all North American Railcar data



**Step 1:** Railinc sends 6-10 million messages per day

**Step 2:** Refine the messages each day

- Fill table with data into columns
- Flag duplicate messages
- Convert timestamps to CST
- Flag new messages

**Step 3:** Make messages available for the Finance Team

- Add messages to the Main Data Table
- Finance Team queries from the Main Data Table to find missing billings

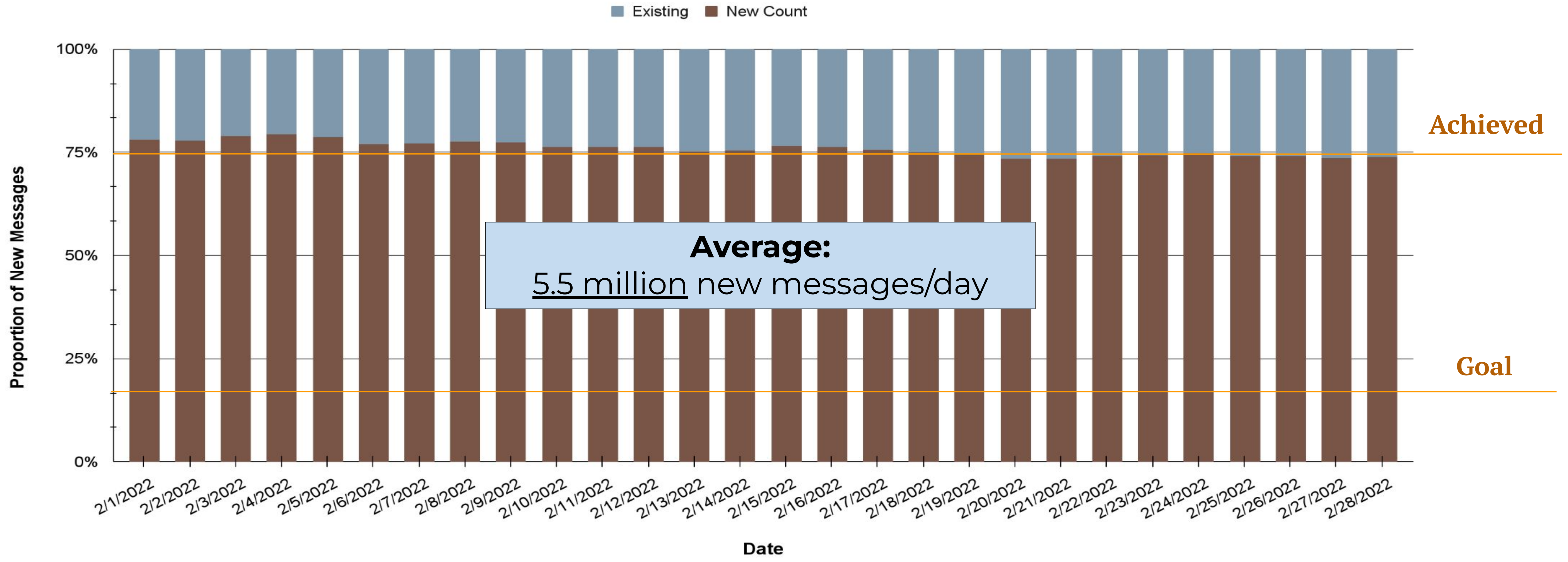
## ETL Data Pipeline Solution:

- Adds missing (offline) data to the Main Data Table
- Extracts, Transforms, and Loads (ETL) raw daily messages from Railinc
  - Flag the duplicate messages
  - Fix the incorrect city and state names
  - Convert the message's local time to central timestamp
  - Find any matches between Railinc messages and messages from other data sources
- Finance Team can query from the Main Data Table

## Results

- Initially, the Finance team wanted at least 20% of the messages that we pipelined into the main data table each day to be new messages
- However, roughly 75% of the messages we are pipelining in to the main data table are new messages
- Each day, we provide the finance team with about 5.5 million new messages that can help them find missing billings

***Outcome: Exceeded Finance Team's Requirement of 20% New Messages Everyday***



**Average:**
5.5 million new messages/day