Data Analytics Workshop

Case #1 – Factory Workers Air Quality Index

Author: Nikolay Shopov
Role: Managing Director
Email / Phone: nikolay.shopov@adastragrp.com / +66 66 059 7492
Office: 24088, 24 Fl, Chamchuri Square, 319 Phaya Thai Road, Pathum Wan, 10330 Bangkok

Company Name: Adastra (Thailand) Co., Ltd
Company ID: 0105561192034

Date: 11-Nov-2019

1. Adastra Quick Facts

1. Adastra is a group of companies, covering the full spectrum of modern Information Management technology field. Check out our group website to get a grasp: https://adastragrp.com/

2. Adastra specializes in Information Management riding the wave of all those trending words nowadays – Cloud Computing, Big Data, Data Science and AI, Data Warehousing and Business Intelligence at scale

3. Adastra is headquartered in Canada and Czech Republic but we are Adastra (Thailand) Co., Ltd. and we are proud to call Bangkok our home

4. Adastra was founded 20 years ago but moved to Thailand just recently in 2019 as a BOI promoted high value-added digital company. We like to consider ourselves as a fast-moving international company with the spirit of a startup and the expertise and resources of a large multimillion corporation

5. We speak 23 languages natively in our 23 offices in North America, Europe, Asia and Australia. We hit the 2,000 employees mark having completed 400+ projects for large and small clients within the banking, telco, finance, retail, insurance, automotive and public sectors. We have 32 international awards for quality and have been recognized as a Platinum Member of the Canada Best Managed Companies Award. Proud partners of most the leaders in Information Management technologies.

2. Adastra Students Offerings

1. Adastra Academy
- Take extra subjects in Data Technologies under the mentorship of Adastra Thailand experts
- Study from Home/Uni at your own pace, all for free
- Estimated Duration – 3-6 months
- Graduates will be admitted to work at Adastra Thailand, with a premium sign-on bonus

2. Internship Onsite
- Utilize the Adastra Academy content onsite in our offices
- Estimated Duration - 2 months

- Graduates will be admitted to work at Adastra Thailand, with a premium sign-on bonus

3. Open Junior Positions
- We offer full time junior positions
- Flexible Hours - Work from Home/Uni, Part Time (4 hour), Evenings/Mornings Work
- Graduates from this workshop – No Technical Interview!

2. Data Case Overview

An industrial company in Europe has been struggling in the past with worsening work related conditions in their factories. Several incidents were recorded with employees suffering from lung diseases that were tied to polluted factory environment. Workers labor unions threatened with protests and absence from work if our client does not install stations to measure the quality of air in the factories. An investment was undertaken to address the problem and sensors were installed to track the air quality in the factories. We have been engaged to measure the impact of this investment, understand the current situation and alert in case of future required investments.

3. Source Data Description

The source file includes the following columns:
1. **[timest] = Timestamp of the Sensor Recording, Timestamp**
   a. Format=[YYYY:MM:DD HH:MM:SS]

2. **[station] = ID of the factory measuring station, Integer**
   a. 1=Factory Mercury
   b. 2=Factory Venus
   c. 3=Factory Earth
   d. 4=Factory Mars
   e. 5=Factory Jupiter
   f. 6=Factory Saturn

3. **[param] = ID of the measured parameters, Integer**
   a. 0=PM
   b. 1=NO2
   c. 2=NO
   d. 3=C6H6
   e. 4=CO
   f. 5=O3
   g. 6=SO2
   h. 7=Humidity
   i. 8=Atmosphere Pressure
   j. 9=Wind
   k. 10=Sunshine
   l. 11=Temperature

4. **[level] = Sensor Reading for corresponding measure, Decimal**

5. **[isinvalid] = Sensor Error Indicator, Integer**
    a. 1=Error
    b. 0=Pass

4. Project Deliverables

Client is asking for a set of deliverables. Feel free to address the project deliverables as a team. Partial solutions are welcome as long as they address at least 1 out of the 3 problems below:

4.1. AQI as a Single Comparable Measure Across Factories

Having a set of measures does not make things easy for comparison.

The Air Quality Index (AQI) is a standard single measure for evaluation of the air quality based on a number of metrics. You can read about it on Wikipedia (Link) or other public sources. While there are multiple ways to calculate it, our client is using the NowCast method, explained well here: Link.

4.2. Descriptive Analytics - Standard Visualizations to Describe Current State

Client is asking for a number of visualizations that they can use to oversee the quality of air in the factories at any point of time. What are the measures and trends that you deem most important and how would you visualize them best?

Examples:
1. How many days per month (in average) do we observe dangerous levels of air pollution?
2. What's the longest streak of days with dangerous levels?
3. How is weather correlated to air pollution in the open-space factories?
4. What's the hour window (00-06, 06-12, 12-18, 18-24) with highest air pollution?
5. What's the factory with best and worst air conditions?

4.3. Predictive Analytics – Predict Air Quality for Next 6 Months

Our client is under statutory regulations and must ensure that the air quality falls into permissive range of [Good] or [Moderate], i.e. AQI<=150 at all times. Based on historical data what model would you use to predict the AQI for the next 6 months (starting from last available data point) in case our client does not take any action today?

5. Data Driven Approach

0. Set of Tools
- You are free to use any set of tools, programming languages or libraries you wish
- On our side we may recommend Anaconda (with python/pandas/matplotlib) and free desktop version of BI tools such as MS PowerBI, Microstrategy or Tableau.

1. Data Acquisition & Profiling
- Load the data into your profiling tool of choice
- Get to know the dataset – simple statistics, distinct values, data types, etc.

2. Data Cleansing
- Check and convert column data types
- Special symbols processing
- Add/remove features (columns) that will (not) be used later
- Rename columns to meaningful names
- Capture, analyze and remove outliers in data
- Capture bad data
- Fill in missing values
- Others

3. Data Modelling
- In a scenario of multiple files create a relational model to work with facts and dimensions (N/A)
- In a scenario with a single composite file pivot rows to columns to match your needs

4. Analytics
- Use appropriate charts to visualize your findings. Combine some in a mockup dashboard
- Choose appropriate machine learning model to predict future values based on available data

6. Next Steps

1. Workshop
- Split into teams of max 5 students
- Address the problems by drafting a technical plan
- Use limited coding to support your ideas, but full solution is not expected within the workshop

2. Group Project Work at Home or Class
- Continue working on the full scope implementation from home or university class
- For support, reach out to: Ruj.Akavipat@adastragrp.com
- Deliver your solution to HRTH@adastragrp.com no later than the end of the school year
- Best solutions qualify for a no-technical-interview admission at a Junior position and career at Adastra Thailand and up to **30,000 THB** sign-on bonus