Project Proposal CDL IoT Use Cases

Team members: Yi Chen, Henry Liang, Sharika Mahadevan, Ruben Nakano, Samuel Swain, Yumeng Zhang

Problem Statement

REFiT is a one-of-a-kind cloud service that allows businesses to harness the power of Internet of Things (IoT). It captures streaming data from these devices in real-time, augments it with static information to make it meaningful and generates business values by making predictions that could address the woes of the industry. Therefore, it provides an integration layer between the services that produce events and systems that consume these events, such as Grafana or Apache Cassandra.

REFIT differentiates itself from the other available options in the market by deploying Northwestern’s ML research with flexibility, building a foundation on latest open-source tools and most importantly, through the ease of scalability of ML models.

In the past, REFIT has implemented multiple use cases across industries such as

- **Agriculture**: Based on sensors used in the field, REFIT can enable study of preventive maintenance or controller release of pesticides/fertilizers needed for growing crops.
- **Healthcare**: Based on real-time monitoring of vital signs of patients, REFIT model has predictions of development of certain medical conditions.
- **Manufacturing**: REFIT can be used to monitor the status of shipments, equipment, and market conditions taking preemptive actions based on predictive solutions.

Problem Statement for the team is to:

1. Develop and implement 3 use cases based on public data in REFIT. Each use case must be based on streaming data (but it does not necessarily have to be IoT).
2. Implement the same use cases by using one of the three big cloud providers. The specific choice is up to the team.
3. Design and implement a feature selection and engineering algorithm solely relying on time series data.
4. Assess the pros and cons of using REFIT vs Cloud platforms.

Purpose and Objectives

Create three use cases applications for public data collected in the past, completed end-to-end from data ingestion to modeling, prediction and visualization in both popular big cloud providers and REFiT. Have a better understanding of how REFiT works differs from common big cloud providers and then compare the pros and cons of REFiT as well as big cloud providers.

Dataset Selection and Preprocessing:

- Explore multiple historical time series datasets which based on IoT and choose the best three and come up with related use cases
- Clean the data to remove blank and null values and ensure all data is valid and unique
- Generate diagnose plots to detect and remove potential outliers and influential points

Use cases application:

- Script to read historical data to stream data in to REFiT solution
- Design and implement a feature selection and engineering algorithm solely relying on time series data
- Simulate streaming using historical data and perform predictions using this simulated streaming data as well as store this data the fetch to make the latest and most accurate predictions
- Choose the best machine learning models to perform predictions

Comparison between REFiT and Big cloud providers:

- Choose the most suitable cloud platform to apply same use cases
- Compare the performance of REFiT and the cloud platform for each use case based on quantitative results
- Assess pros and cons of using REFiT vs cloud platforms
- Provide suggestion for choosing REFiT or cloud platform based on the type of the datasets and goals of projects

Project Deliverables

The NU student group will be responsible for completing several deliverables for CDL through the practicum project.

1. **Deliverable 1**: The first deliverable will be a holistic project report detailing the findings in practicuum. These findings will provide the foundation for accomplishing the objectives of this project and will help address the challenges the CDL is facing. For instance, it will provide a detailed analysis of performance of REFIT vs current market leader (potential cloud provider: AWS).

2. **Deliverable 2**: Implementation of the 3 specific use cases of streaming data.

The information contained in the final report will be presented to the following key project stakeholders: Diego Klabjan (Technical Advisor) and Borchuluun Yadamsuren (Business POC) by the end of the project.

Tools and Methods

We will need multiple tools in combination with machine learning models/methods to tackle multiple use cases. AWS will be the backbone of our project. It provides on demand cloud computing for any company, individual, or organization that needs it. It will supply the computing power to train, test, and deploy models. Since AWS charges as a user consumes compute power, we will need funding for this part of the project.

REFiT is one of the two methods our team will be using to predict outcomes based on data. It is a system built to quickly and efficiently process IoT streaming data from the billions of IoT devices connected to the internet. We will be sending simulated data streams straight into REFiT to receive predictions on interesting dependent variables. It was designed to be used with major cloud providers such as AWS.

We will be using Python to do many things such as clean, explore, and understand the data we find for our project. Python is a general purpose programming language that excels when it comes to data science. One package specifically within Python, Pandas, allows us to wrangle the data into any form we could possibly need. Pandas is so powerful because it offers the data frame and sequence structure that can be operated on to result in any number of tables needed for analysis.

Scala will also be used for this project. REFiT is built on Scala and Python. We will need both programming languages to understand the REFiT platform. Scala, similar to python, supports object oriented programming and functional programming. For this reason we hope to quickly learn Scala.

Lastly, we plan on using SQL as our database management system. SQL is designed to work with large amounts of data quickly and efficiently. Using Pandas with large amounts of data can result in long waiting times or even memory errors. We plan to do our data selection in SQL to avoid this.

As we keep learning more in and out of class, we plan to keep developing new and better methods to approach this problem. Thus far, we know we will be evaluating the pros and cons of many different machine learning models and data analysis methods to come up with the best result. The specifics of which can’t be known until we find the datasets and problems we will be working with.

Scope: define responsibilities

The NU student group will complete the project deliverables detailed in that section. Any item not listed will not be required to be completed.

The NU student group will work to look for, review, analyze, and clean time series data from public data sources. Private data will not be considered in this project, and the NU student group will provide CDL with data cleansing documentation upon request, but will not be included in the final report by default.

The NU student group will use three of such public data sources to develop three distinct use cases of the data. The student group will simulate the streaming of said data such that it mimics data streaming in real time from a device in an Internet of Things (IoT) system.

The NU student group will develop and implement feature selection and engineering algorithms based on said time series data. The student group will also use said time series data to do machine learning and/or inference. Feature selection, feature engineering, and inference will be integrated into REFIT, simulating the complete workflow of REFIT in each of the selected use cases, including visualizations. The same three use cases will also be developed using cloud computing, including data ingestion, feature selection and engineering, inference, and visualization.

The NU student group will not perform a review or conduct any change to the existing REFIT architecture. The student group will use open source technologies and other technologies made available by CDL. The NU student group will have the final decision on the technologies that will be used to build the use cases in REFIT and cloud platform unless explicitly required by CDL. Suggestions from CDL are welcomed and encouraged, but lack thereof will not impact success of this project.

Risks and Limitations

1. Current REFiT prototype has issues with version compatibility and is undergoing bug fixing. If not fixed in time, it will delay the implementation of use cases on REFiT.
2. The lack of documentation on REFiT and resources who understand the system makes it more difficult to implement use cases on REFiT.
3. Further analysis on data for each use case may show that some of them are not suitable for this project. Therefore, all use cases referenced in this document are not final.
4. Because of these limitations milestones are subject to change.

Potential Use Cases (subject to data sufficiency)

1. Divvy bikes
    a. Use Divvy bike historical data to simulate real time information on trips and bike availability in each dock
    b. Model future demand

   c. Use external data such as temperature, weather and important dates (holidays, game days, etc) to make predictions more accurate and flexible
   d. Data available at: https://ride.divvybikes.com/system-data
2. Stock Market prediction (If hourly stock price data is available)
3. To be determined

Milestones

- **FALL QUARTER (10/02/2022 - 01/01/2023)**
    - **Understand practicum project scope and write business proposal (10/02/2022 - 10/30/2022): 4 weeks**
        - Clarify the scope and objectives of this practicum
        - Explore multiple public time series data and select best 3 datasets for use cases applications
        - Complete business proposals
    - **Platform set-up and finalize use case selection (10/30/2022 - 11/12/2022): 2 weeks**
        - Finalize use case selection
        - Learn about REFiT by its documentation and set up the environment on personal computers
        - Research on common big cloud providers about their functions and select one to work on
        - Learn and identify cloud tools to be used
    - **EDA on first use case (10/30/2022 - 11/19/2022): 3 weeks**
        - Clean and explore the first use case dataset and deliver EDA summary
        - Research modeling methods for first use case
    - **Implementation of the first use case (11/13/2023 - 12/10/2023): 4 weeks**
        - For REFiT and cloud provider
            - First complete implementation (working code and all pieces put together)

- **WINTER QUARTER (01/02/2023 - 03/28/2023)**
    - **Completion of the first use case (01/02/2023 - 01/26/2023): 4 weeks**
        - For REFiT and cloud provider
            - First complete implementation (working code and all pieces put together)
            - Testing and debugging
            - Evaluation
        - Present our findings on the first use case
        - Complete the comparison of the first use case between cloud providers and REFiT
        - Write section on first use case

- **Completion of the second use case (01/29/2023 - 02/26/2023): 7 weeks**
    - Clean and explore the second use case dataset and deliver EDA summary
    - Research modeling methods for second use case
    - For REFiT and cloud provider
        - First complete implementation (working code and all pieces put together)
        - Testing and debugging
        - Evaluation
    - Present our findings on the second use case
    - Complete the comparison of the second use case between cloud providers and REFiT and report our findings
    - Write section on second use case
- **Implementation of the third use case (03/19/2023 - 03/28/2023): 2 weeks**
    - For REFiT and cloud provider
        - First complete implementation (working code and all pieces put together)

- **SPRING QUARTER (03/28/2023 - 06/03/2023)**
    - **Completion of the third use case (03/28/2023 - 05/04/2023): 6 weeks**
        - Clean and explore the third use case dataset and deliver EDA summary
        - Research modeling methods for third use case
        - For REFiT and cloud provider
            - First complete implementation (working code and all pieces put together)
            - Testing and debugging
            - Evaluation
        - Present our findings on the third use case
        - Complete the comparison of the second use case between cloud providers and REFiT and third our findings
        - Write section on third use case
    - **Finish the final report (05/07/2023 - 06/10/2023): 5 weeks**
        - Combine findings from three use cases together and finish final report with overall conclusion
        - Debug and implement changes on previous use cases (if needed)

Link to detailed timeline:
https://docs.google.com/spreadsheets/d/1S76sHJwkBTZ11L0fU1f6ssYZIDGjZjQSZAFq4HTC6J0/edit#gid=0

	Quarter	FALL													WINTER												SPRING										
	Month	10	10	10	10	10	11	11	11	11	12	12	12	12	1	1	1	1	1	2	2	2	2	3	3	3	3	4	4	4	4	4	5	5	5	5	6
		10/2/2022	10/9/2022	10/16/2022	10/23/2022	10/30/2022	11/6/2022	11/13/2022	11/20/2022	11/27/2022	12/4/2022	12/11/2022	12/18/2022	12/25/2022	1/1/2023	1/8/2023	1/15/2023	1/22/2023	1/29/2023	2/5/2023	2/12/2023	2/19/2023	2/26/2023	3/5/2023	3/12/2023	3/19/2023	3/26/2023	4/2/2023	4/9/2023	4/16/2023	4/23/2023	4/30/2023	5/7/2023	5/14/2023	5/21/2023	5/28/2023	6/4/2023
Understand scope and write proposal		■	■	■	■																																
Platform set-up						■	■																														
Finalize use cases						■	■																														
FIRST USE CASE																																					
	EDA					■	■	■																													
	Research modeling methods					■	■	■																													
	Complete implementation							■	■	■	■				■	■																					
	Testing and debugging																■	■																			
	Evaluation																■	■																			
	Compare REFiT and cloud provider																	■																			
	Present findings																	■																			
	Write report section																	■																			
SECOND USE CASE																																					
	EDA																		■																		
	Research modeling methods																		■																		
	Complete implementation																			■	■	■	■														
	Testing and debugging																							■	■												
	Evaluation																							■	■												
	Compare REFiT and cloud provider																								■												
	Present finding																								■												
	Write report section																								■												
THIRD USE CASE																																					
	EDA																									■											
	Research modeling methods																									■											
	Complete implementation																										■	■	■	■							
	Testing and debugging																														■	■					
	Evaluation																														■	■					
	Compare REFiT and cloud provider																															■					
	Present finding																															■					
	Write report section																															■					
FINAL REPORT																																	■	■	■	■	■

Signatures

Henry Liang:

Sam Swain

Yi Chen

Yumeng Zhang

Shaika Mahadevan

Ruben Nakano