ANZ Hybrid Cloud Analytics

Mentor: Alan Litchfield **Moderator:** Kenneth Johnson **Client:** ANZ
Team Member: Carl Jason Chang, Yoonkun Lee, Yinjia Lee, Hohua Diamond

AUT

ABOUT THE PROJECT

BACKGROUND

ANZ is an Australian multinational banking and financial services company headquartered in Melbourne and are known as one of the largest banks in New Zealand.

Currently, ANZ's existing internal cloud is an Open-Shift cluster environment run on internal hardware with limited computing resources. If an attempt was made to process ANZ's reservoir of big data, the system will crash.

Due to the sensitivity of the data that ANZ stores (customer information), deployment of customer data in its raw form to any outsourced/offshore services is prohibited by law (BS11, RESERVE BANK OF NEW ZEALAND, 2020). This raises the problem where data cannot freely move from ANZ's OpenShift cluster to GCP.

This project is to provide a proof of concept for the integration of a hybrid solutions to ANZ's Machine Learning Pipeline(MLP).

RATIONALE

For this project, ANZ will use the ML pipeline as proof of concept for advanced data analytics and machine learning with hybrid cloud (a connection between internal cluster and cloud services) solutions.

USE-CASE

Given the use-case, how might we predict the number of houses / home loans in Wellington so that the bank can prepare adequately for demand? Build a model with the pipeline.

OBJECTIVES

	ACHIEVED
Provide an architecture diagram for the solution.	☑
Provide recommendations, and pros and cons of what you have found.	☑
Create an OpenShift cluster.	☑
Client to create sensitive data for the cloud.	☑
Create a secure method to deploy sensitive ANZ data on the cloud.	☑
Source and join public and (synthetic) sensitive private data.	☑
Train and deploy machine learning model(s) that solve the proposed use-case.	☑
Create a secure model management process.	☑
Create a dashboard to communicate your data insights (i.e. visualisation).	☑

ARTIFACTS & PROCESS

P1: BUILD THE OPENSHIFT CLUSTER

Main artifacts produced

-OpenShift Origin Platform v3.7.2 (Red hat, 2020)
OpenShift is used to manage applications and deploy pods to different nodes. In this project the openshift v3.7.2 is built for managing the jupyter notebook application and MySQL server.

-MySQL Databases v5.7 managed by OpenShift
MySQL server is considered as an essential part in this project to save all raw data. Therefore, the MySQL Database was created and deployed.

-Jupyter notebook v3.6 managed by OpenShift
An open-source web application that allows users to create and share documents that contain live code, equations, visualizations and explanatory text. In this project, the Jupyter notebook v3.6 is imported into OpenShift for the data preparation stage.

P2: DATA PREPARATION (JUPYTER NOTEBOOKS)

-Query: Retrieve data from an internal database on premise.

-Clean: Correcting incomplete, incorrect or irrelevant parts of the data

-Transform (encoding and standardization) (Grus, 2015): One-hot encoding on categorical variables. Feature scaling - rescaling data to have mean 0 and standard deviation 1.

-Upload: Transfer the cleaned dataset to the data lake (BigQuery)

P3: DATA LAKE (BIGQUERY)

-Business Case Zone: Stores data relevant to the business case

-External landing Zone: Raw data sourced from external resource

-Clean Zone: Cleaned versions of data sets

-Internal landing Zone: Data sourced from internal on-premise resources

-Model: Stores models

-Model eval: Stores evaluation metrics about the models

P4: MODELLING (BIGQUERY ML)

-Train: Uses the prepared dataset in the business case zone to train the model.

-Evaluate: Evaluates model metrics

-Test: Test the model with unseen data

-Predict: Prediction on the most recent data set.

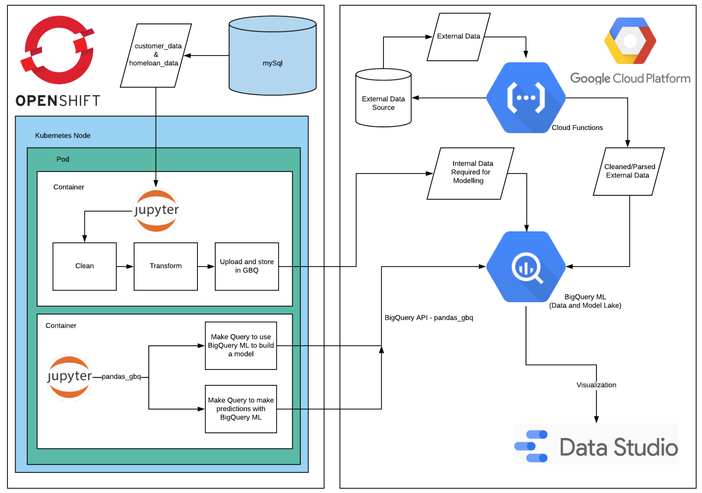

Fig 1. System Architecture

P5: DASHBOARD (GOOGLE DATA STUDIO) (GOOGLE, 2020)

Main Artifacts Produced:
- Growth of new ANZ home loan
- ANZ Market share for new home loan
- Predict the number of home loan
- Average house sold price
- Geographical map that represent the number of house sold

Evaluation of the impacts
- ANZ can predict the upcoming revenue and better manage cashflow.
- ANZ knows the home loan market share so that they can make a strategy and plan.
- ANZ can easily review the trend of customers who got a new home loan and the average house price by suburb.

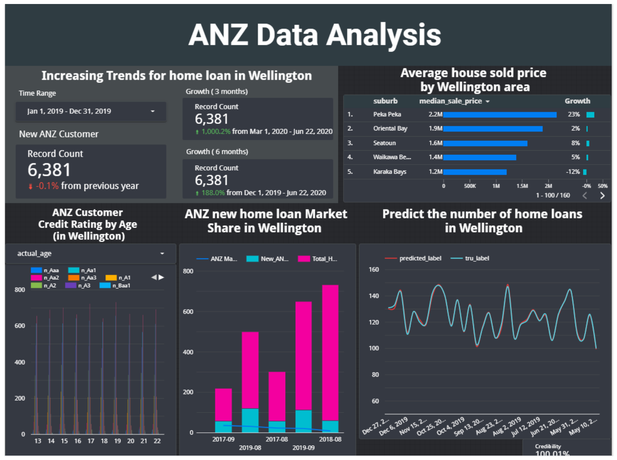

Fig 2. ANZ Data Analysis dashboard

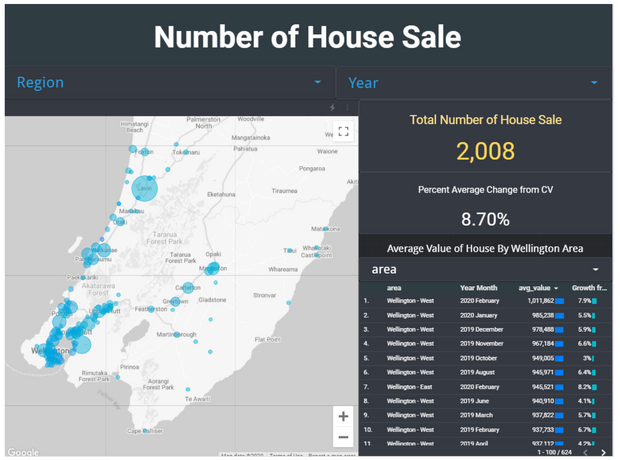

Fig 3. Number of House Sale dashboard

QUALITY ASSURANCE

- Following PEP- 8 (Rossum, Warsaw, & Coghlan, 2001)
- Acceptance Testing for Openshift Cluster and Dashboard (Pietrantuono, Bertolino, De Angelis, Breno, & Russo, 2019)

CHALLENGES

- Each member has a different role that it was hard to support and help.
- The scope was big at the beginning that took lots of time to narrow down.
- This project involves many different areas of knowledge.

FURTHER DEVELOPMENT

- Research Kubeflow pattern
- Get automate public data (scraping complex websites for data)
- The dashboard and data only includes Wellington regions. So these are some restrictions to know the whole trend of New Zealand.
- OpenShift Origin could be replaced by OpenShift Enterprise

REFERENCES • Google. (2020, 06 20). Introduction to BigQuery ML. Retrieved from Google Cloud: https://cloud.google.com/bigquery-ml/docs/bigqueryml-intro • Google. (2020, 06 21). Visualizing BigQuery data using Data Studio. Retrieved from Google cloud: https://cloud.google.com/bigquery/docs/visualize-data-studio • Grus, J. (2015). Data Science from Scratch. Sebastopol: O'Reilly. • Pietrantuono, R., Bertolino, A., De Angelis, G., Breno, M., & Russo, S. (2019). Towards Continuous Software Reliability Testing in DevOps. ACM 14th International Workshop on Automation of Software Test (AST), 21-27. doi:DOI 10.1109/AST.2019.000009 • Red hat. (2020, 06 20). Top considerations for building a production-ready AI/ML environment. Retrieved from Red hat: https://www.redhat.com/cms/managed-files/cl-openshift-for-ai-ml-e-book-f24178-202006-en.pdf • RESERVE BANK OF NEW ZEALAND. (2020, April 01). BS11: Outsourcing policy. Retrieved from Reserve bank of New Zealand: https://www.rbnz.govt.nz/-/media/ReserveBank/Files/regulation-and-supervision/banks/banking-supervision-handbook/bs11.pdf?la=en • Rossum, G. v., Warsaw, B., & Coghlan, N. (2001, July 05). PEP 8 -- Style Guide for Python Code. Retrieved from python: https://www.python.org/dev/peps/pep-0008/