SERCO DATA SCIENCE USE CASES POC: SHOWCASE

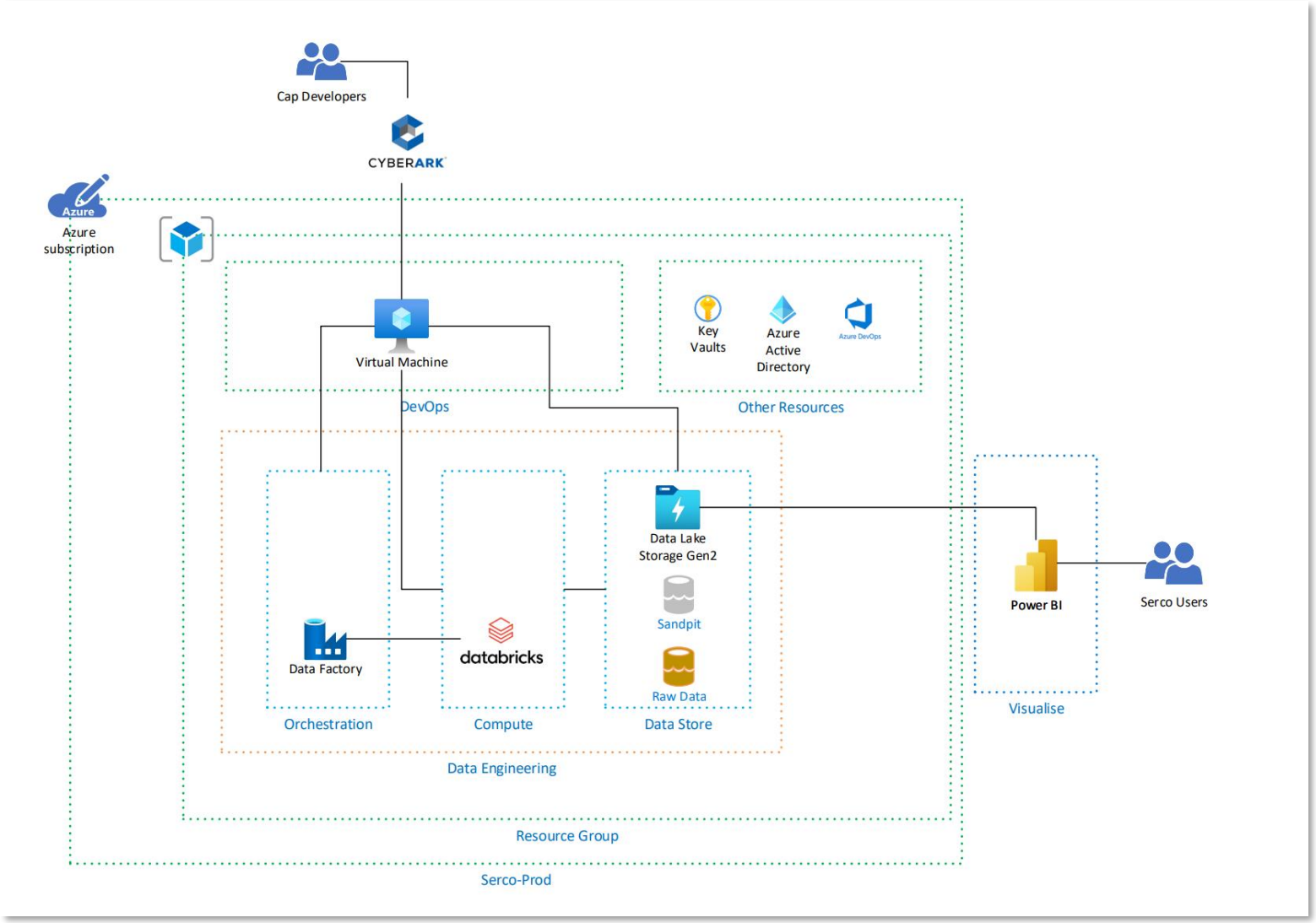
Sprint 3, Week 2

3 February 2023

AGENDA

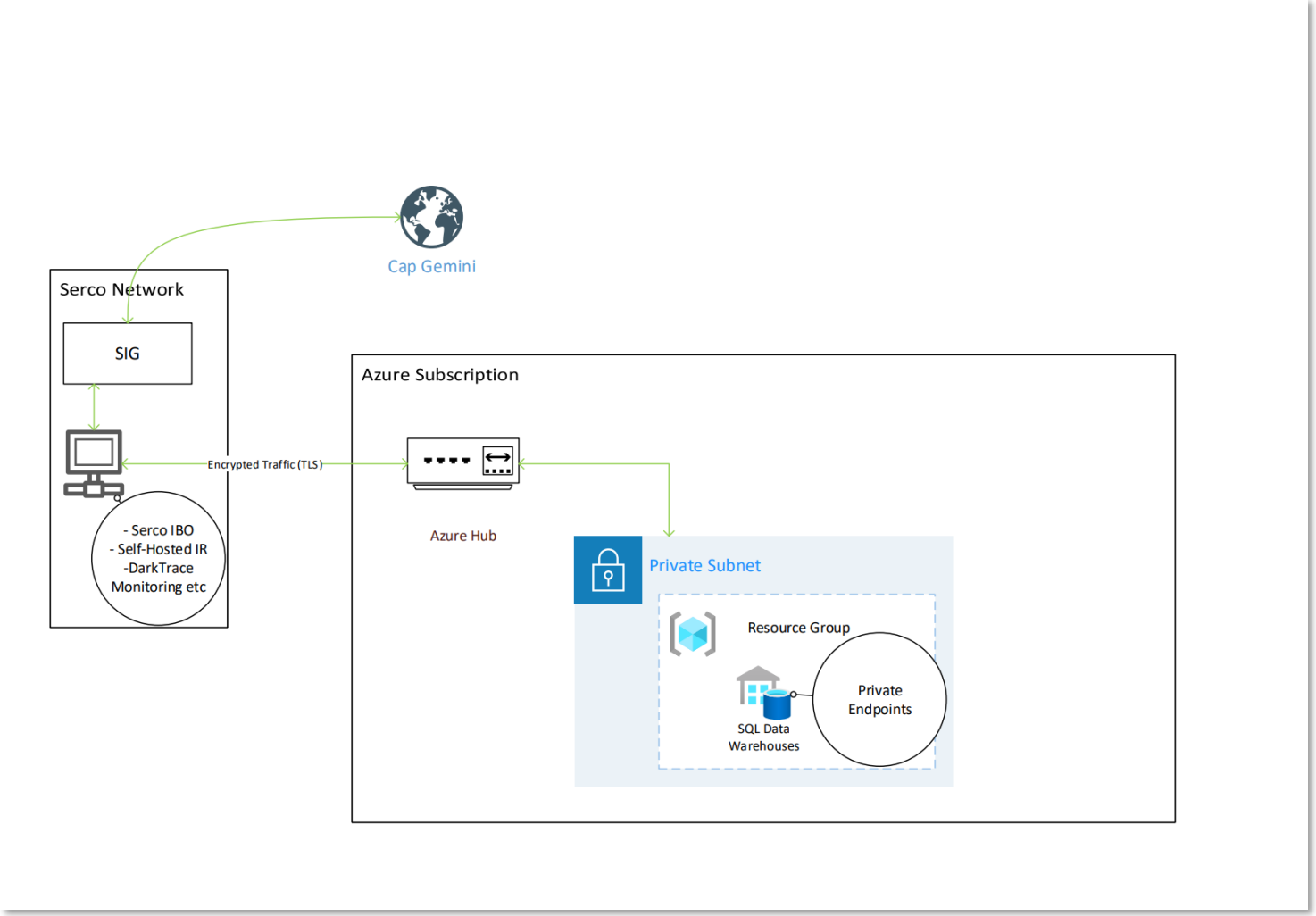
1. Azure data science architecture
2. Network architecture
3. Demonstration of Parquet file connector in Power BI
4. Machine Learning Technique Research
5. Input data model
6. Dashboard visualisation wireframes
7. Azure Tenancy
8. Data Lake Storage
9. UC1 Transport and Escort Demand Model
10. UC2 Incident Forecasting Model

AZURE DATA SCIENCE ARCHITECTURE



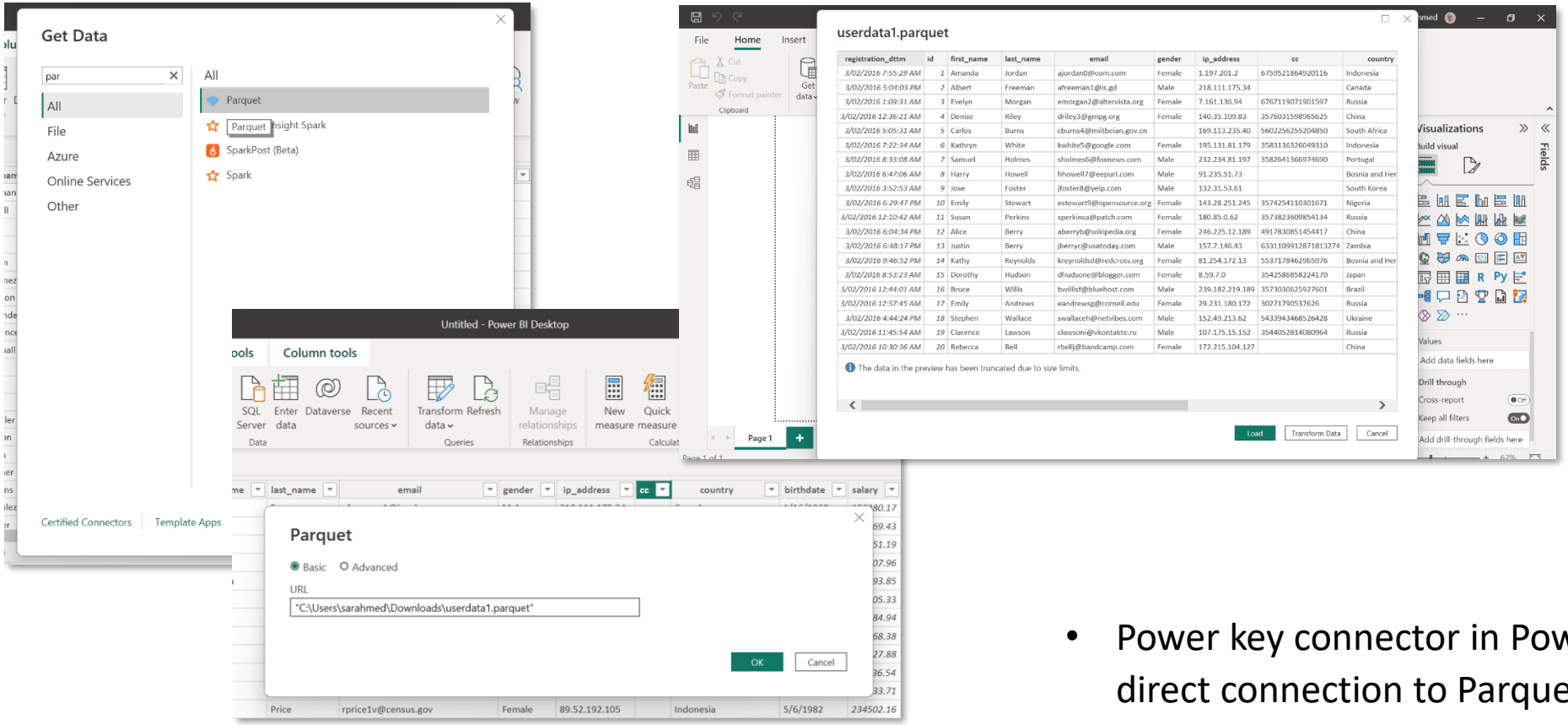
- Azure data science architecture

NETWORK ARCHITECTURE



- Secure remote access for distributed team to Azure tenancy data science environment

DEMONSTRATED PARQUET FILE CONNECTOR IN POWER BI



- Power key connector in Power BI supports direct connection to Parquet files as a source
- Extract the data within your Power BI model.

MACHINE LEARNING TECHNIQUE RESEARCH

Capgemini

SERCO DATA SCIENCE USE CASES POC:
MACHINE LEARNI
RESEARCH

Dr Sara Ahmed
7 December 2022

INTRODUCTION

- Serco Organization recognised the need to enhance the current system to reflect more variables than is currently captured, to better reflect a view of a
needs and responsivity.
- For this reason, designing predictive models that can assess futu
the detainees more accurately than what is already being used is
- Concerned with these issues, the current research attempts to s
the prediction problem in the Serco Organization.
- A number of machine learning techniques were investigated in t
- Including, the logistic regression, random forests, support vector
networks, Search algorithm and Survival Analysis for improving t
- On the other hand , for demand forcasting the techniques Time S
(ARIMA/SARIMA), Regression models, XGBoost, K-Nearest Neigh
Regression(KNN), Random Forest and Long Short-Term Memory
- The performance metrices for the machine learning techniques a

Serco Data Science POC: Discovery Report | 6 December 2022

MACHINE LEARNING TECHNIQUE RESEARCH

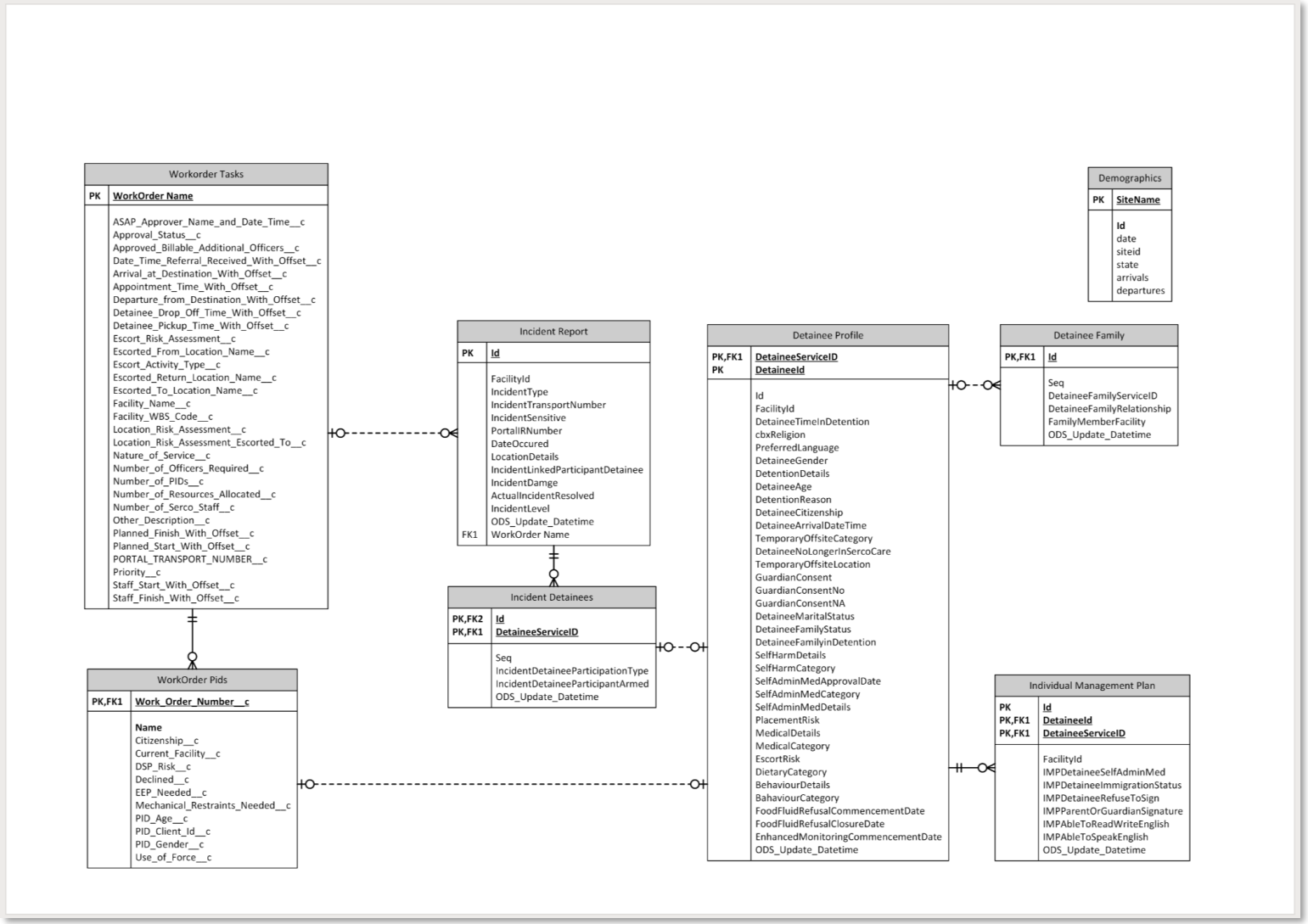
XGBoost / Gradient Boosting and the Search Algorithms

Tool	Description	Potential in predictive performance for risk assessment
XGBoost "Extreme Gradient Boosting"	• XGBoost works by adding a classifier at a time, building upon the previous tree structure, whereas RF randomly and independently trains the classifier. • In other words, the boosted tree model is based on a sequential logic in which each new tree utilizes information gained by the previous trees.	• Since it is relatively new, there have not been published criminological studies that have used this for risk assessment purposes. However, its close relative RF has been used quite extensively.
Search Algorithm	• Search is a software that was designed for searching for structure as a result of a series of research efforts . • Specifically, Search splits the data sequentially on the best binary split of the best variable for reducing the unexplained variance .	• There can be complex combinations that might influence a social outcome. An analyst may not be aware of all possible interactions that should be factored in the modeling process.

Serco Data Science POC: Discovery Report | 6 December 2022

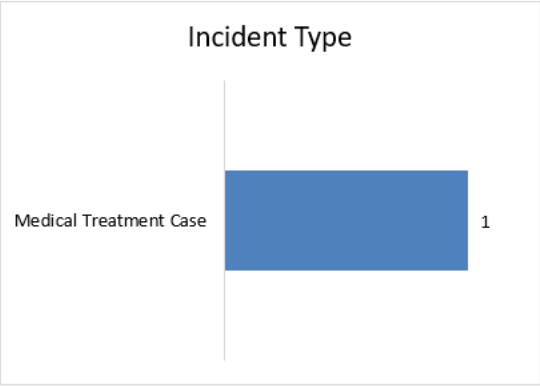
 8

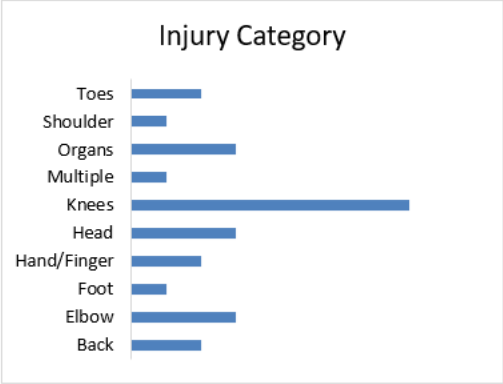
INPUT DATA MODEL

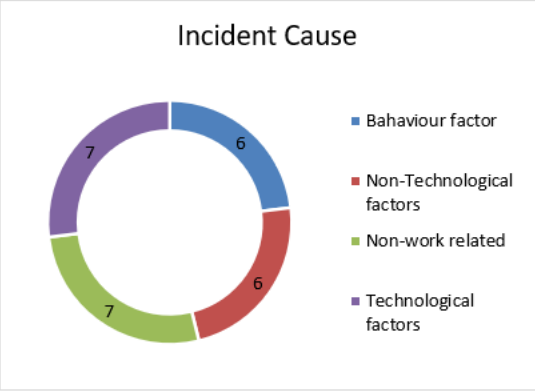


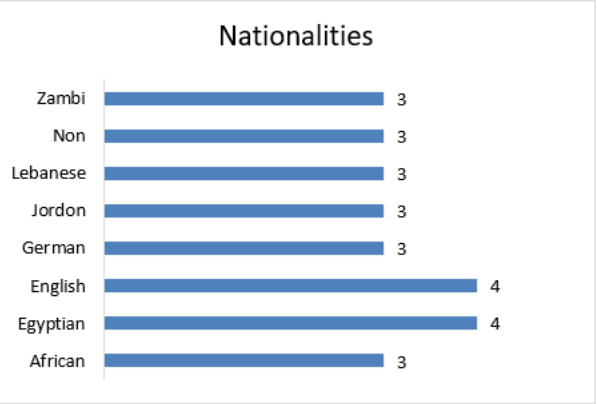
- Data model / entity relationship diagram developed by Chris & Dario can be repurposed for analytics data warehouse

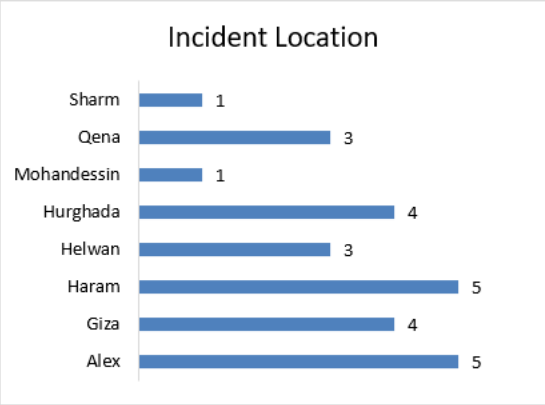
DASHBOARD VISUALISATION WIREFRAMES

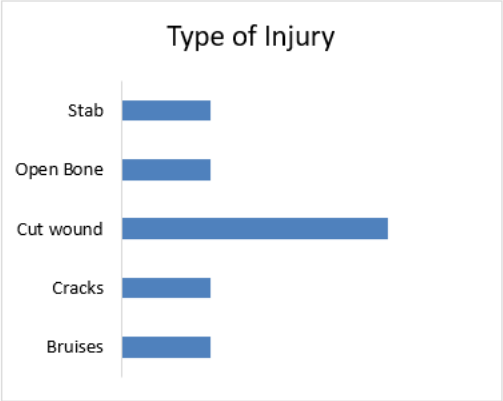


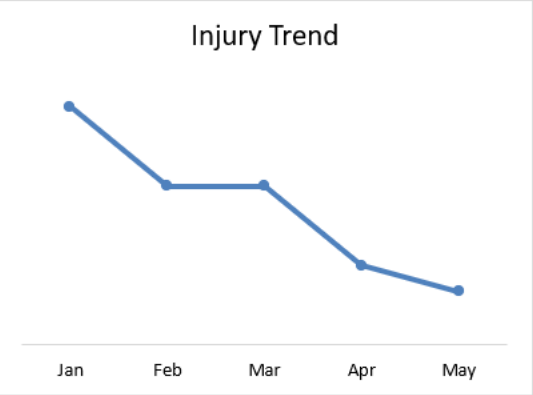


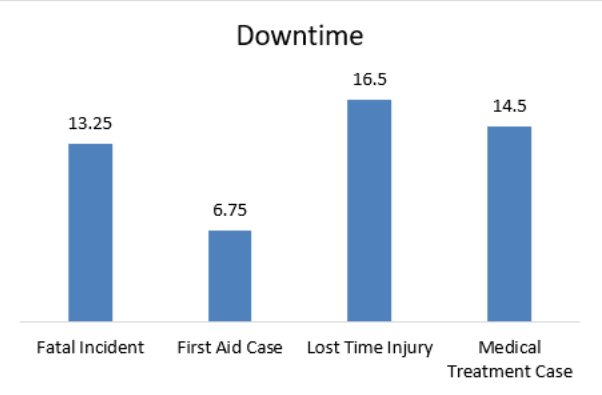












AZURE TENANCY

Azure services

Create a resource | Quickstart Center | Storage accounts | Resource groups | Virtual machines | App Services | SQL databases | Azure Cosmos DB | Kubernetes services | More services

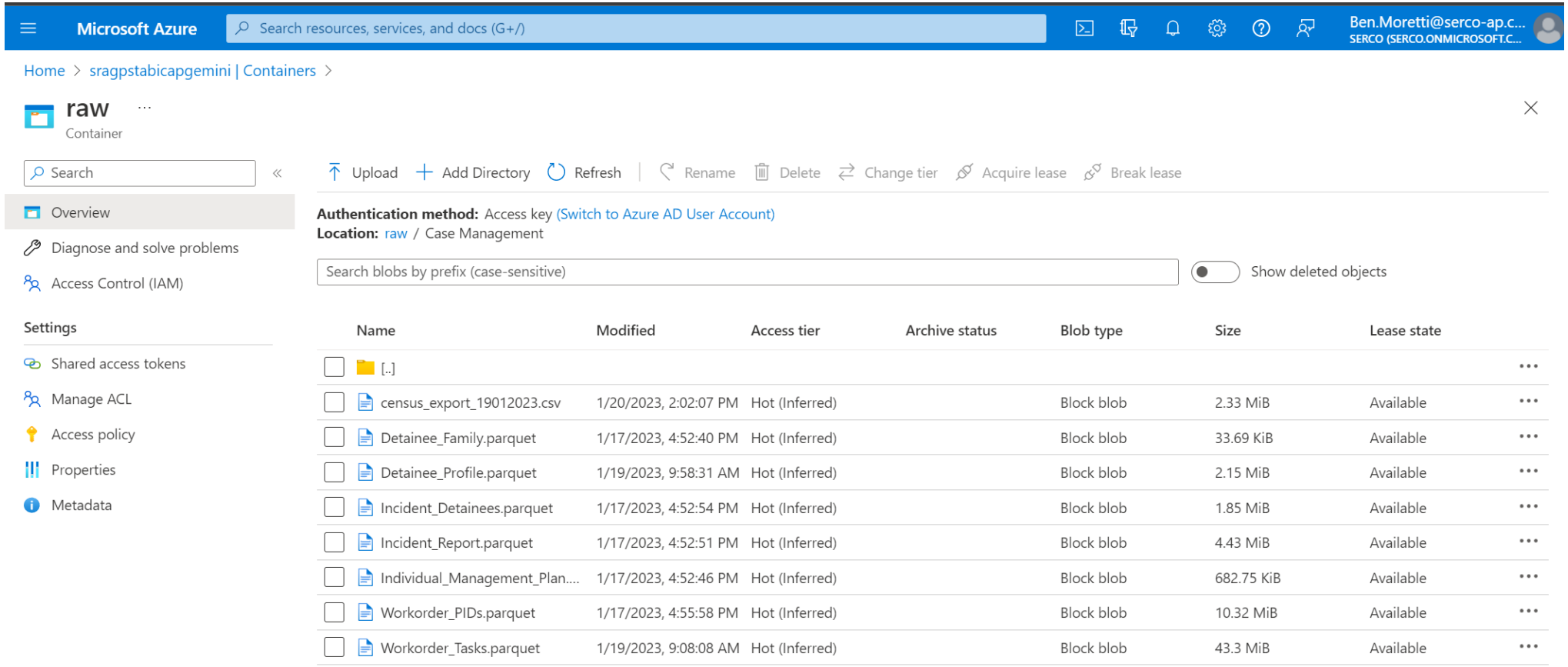
Resources

Recent | Favorite

Name	Type	Last Viewed
SRAGP-ADB-02	Azure Databricks Service	a day ago
sragpstabicapgemini	Storage account	5 days ago
SRAGP-AKV-03	Key vault	2 weeks ago
SRAGP-ADF-03	Data factory (V2)	2 weeks ago
SRAGPI-IDEV02	Virtual machine	3 weeks ago
sragppepbicapgemini	Private endpoint	4 weeks ago
SRAGPI-IDEV02_OsDisk_1_9967c7241416403493456d3949d6b4e8	Disk	4 weeks ago
SRP-ARG-BICapgemini	Resource group	4 weeks ago

See all

DATA LAKE STORAGE



Name	Modified	Access tier	Archive status	Blob type	Size	Lease state
[..]						
census_export_19012023.csv	1/20/2023, 2:02:07 PM	Hot (Inferred)		Block blob	2.33 MiB	Available
Detainee_Family.parquet	1/17/2023, 4:52:40 PM	Hot (Inferred)		Block blob	33.69 KiB	Available
Detainee_Profile.parquet	1/19/2023, 9:58:31 AM	Hot (Inferred)		Block blob	2.15 MiB	Available
Incident_Detainees.parquet	1/17/2023, 4:52:54 PM	Hot (Inferred)		Block blob	1.85 MiB	Available
Incident_Report.parquet	1/17/2023, 4:52:51 PM	Hot (Inferred)		Block blob	4.43 MiB	Available
Individual_Management_Plan....	1/17/2023, 4:52:46 PM	Hot (Inferred)		Block blob	682.75 KiB	Available
Workorder_PIDs.parquet	1/17/2023, 4:55:58 PM	Hot (Inferred)		Block blob	10.32 MiB	Available
Workorder_Tasks.parquet	1/19/2023, 9:08:08 AM	Hot (Inferred)		Block blob	43.3 MiB	Available

UC1 TRANSPORT AND ESCORT DEMAND MODEL

Microsoft Azure | databricks | Search CTRL + P | SRAGP-ADB-02 | ben.moretti@serco-ap.co

UC1-Model-XGBoost-Ben Python

File Edit View Run Help Last edit was yesterday Give feedback

Run all | Terminated | Schedule | Share

Table of contents

- UC1: Predictive demand modellin...
- Configuration
- Data Gathering
- Data Conditioning
- Modelling with XGBoost
- Results Visualisation
- Model Explainability

Cmd 1

UC1: Predictive demand modelling in Transport and Escort to reduce resource costs

Ben Moretti, 30 January 2023

Note: data files have been uploaded locally to DBFS as a workaround until Azure storage container based access can be es

Cmd 2

Markdown

Configuration

Cmd 3

```
# module installation
# install xgboost if it hasn't been done already
#! pip install xgboost
```

Cmd 4

```
import pandas as pd
import matplotlib as mpl
import numpy as np
import pyspark
from pyspark.sql.functions import *
from pyspark.sql import Window
from pyspark.sql.types import IntegerType
# import xgboost related modules
import xgboost as xgb
from xgboost import plot_importance, plot_tree
from sklearn.metrics import mean_squared_error, mean_absolute_error
```

Command took 0.02 seconds -- by ben.moretti@serco-ap.com at 03/02/2023, 1:47:54 pm on SRAGP-ADB-02-Cluster01

Cmd 5

Data Gathering

Cmd 6

```
# read the parquet files into spark dataframes
```

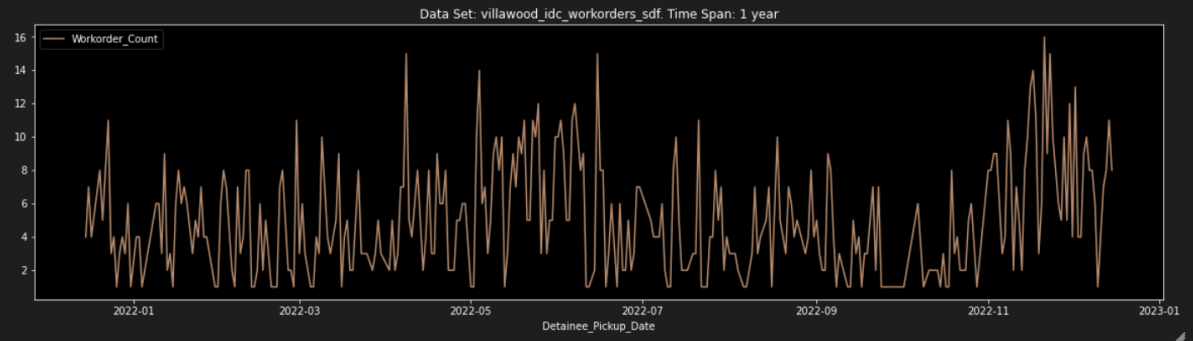
Cmd 23

Python

```
# looks like the data changed significantly after about March 2020 which is obviously due to COVID
# we will subset the data to only provide the last year of data for 2022
villawood_idc_workorders_sdf \
    .select( \
            col('Detainee_Pickup_Date'), \
            col('Workorder_Count'),  \
            datediff(to_date(lit('2022-12-15'), 'yyyy-MM-dd'), col('Detainee_Pickup_Date')).alias('date_diff')
           ) \
    .filter(col('date_diff') <=  365) \
    .drop(col('date_diff')) \
    .toPandas() \
    .set_index('Detainee_Pickup_Date') \
    .plot(figsize=(20,5), title = 'Data Set: villawood_idc_workorders_sdf. Time Span: 1 year')
```

(2) Spark Jobs

Out[12]:

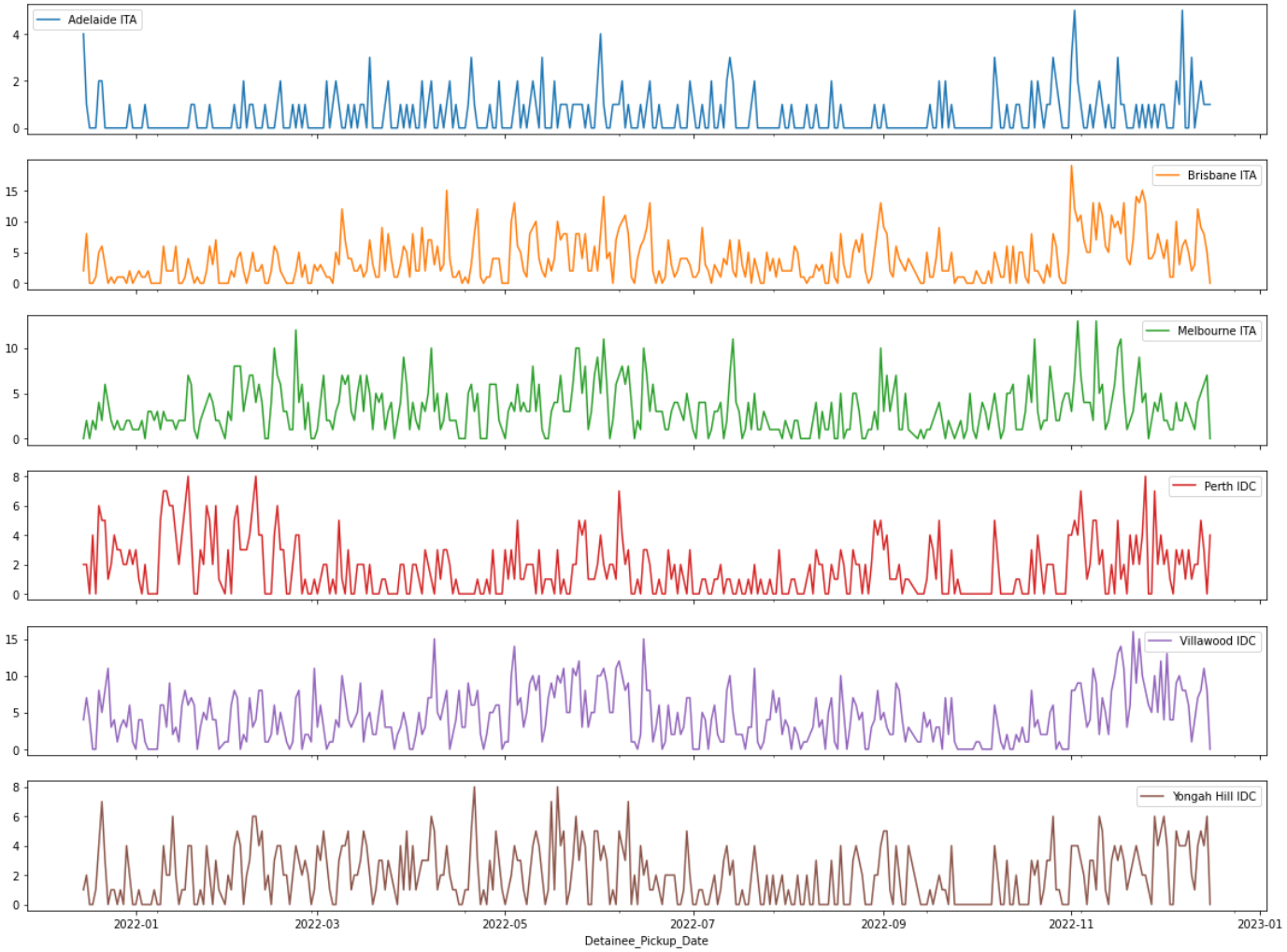



```
<AxesSubplot:title={'center':'Data Set: villawood_idc_workorders_sdf. Time Span: 1 year'}, xlabel='Detainee_Pickup_Date'>
```

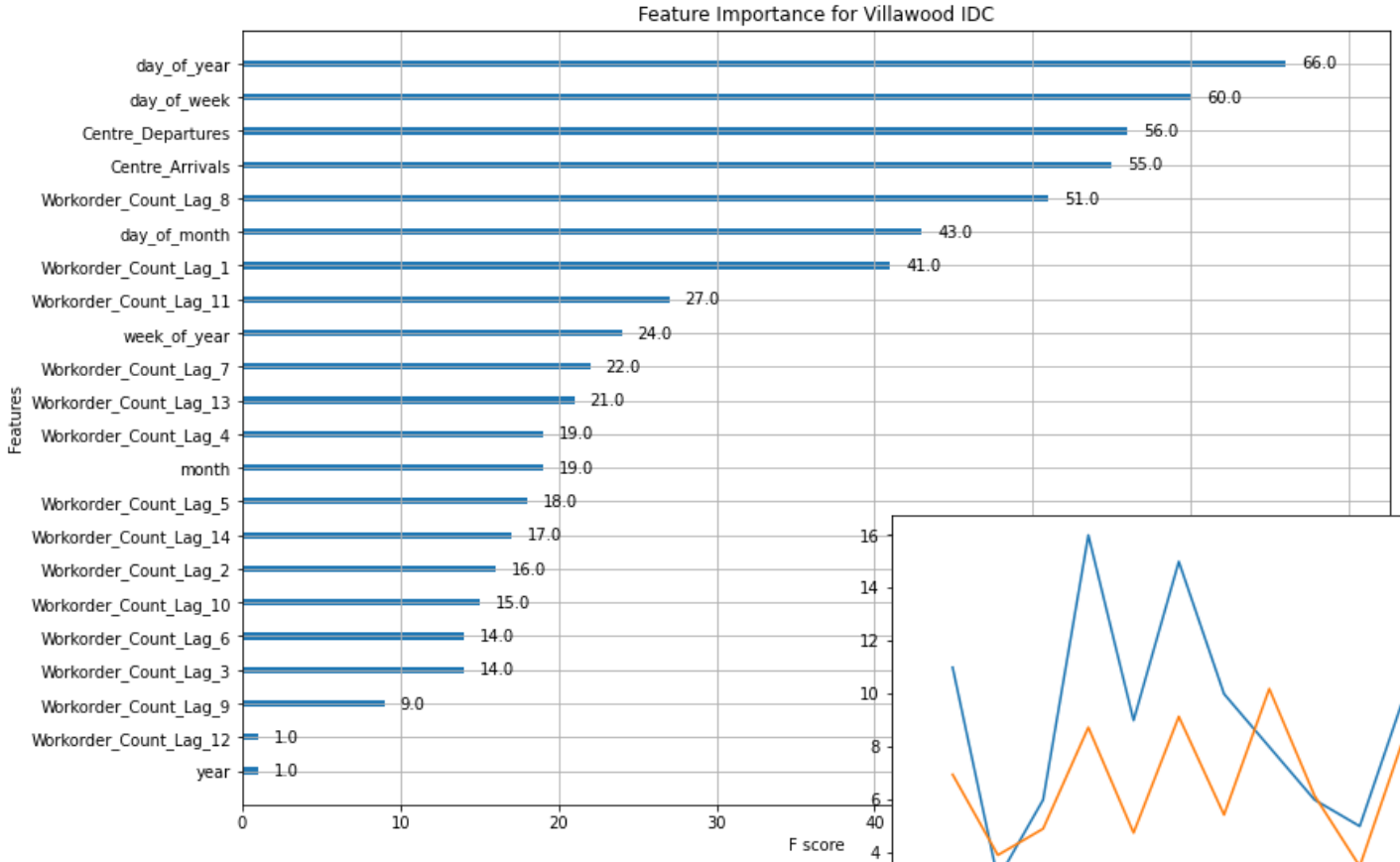
1

Command took 1.34 seconds -- by ben.moretti@serco-ap.com at 03/02/2023, 8:26:11 am on SRAGP-ADB-02-Cluster01

T & E EVENTS OVER TIME



UC1 TRANSPORT AND ESCORT DEMAND MODEL



UC2 INCIDENT FORECASTING MODEL