

SERCO DATA SCIENCE USE CASES POC: SHOWCASE

Sprint 3, Week 2
3 February 2023



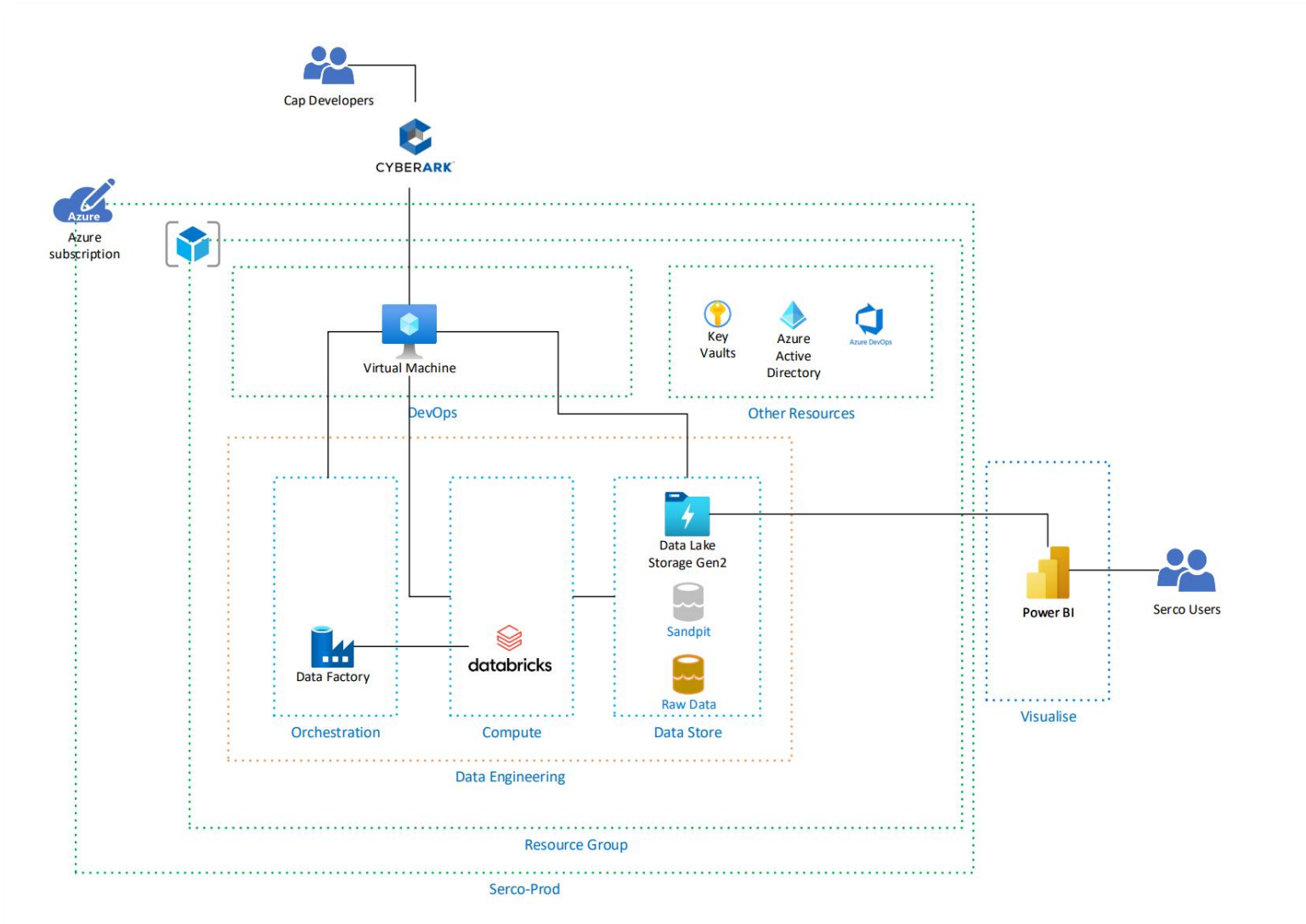
AGENDA

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



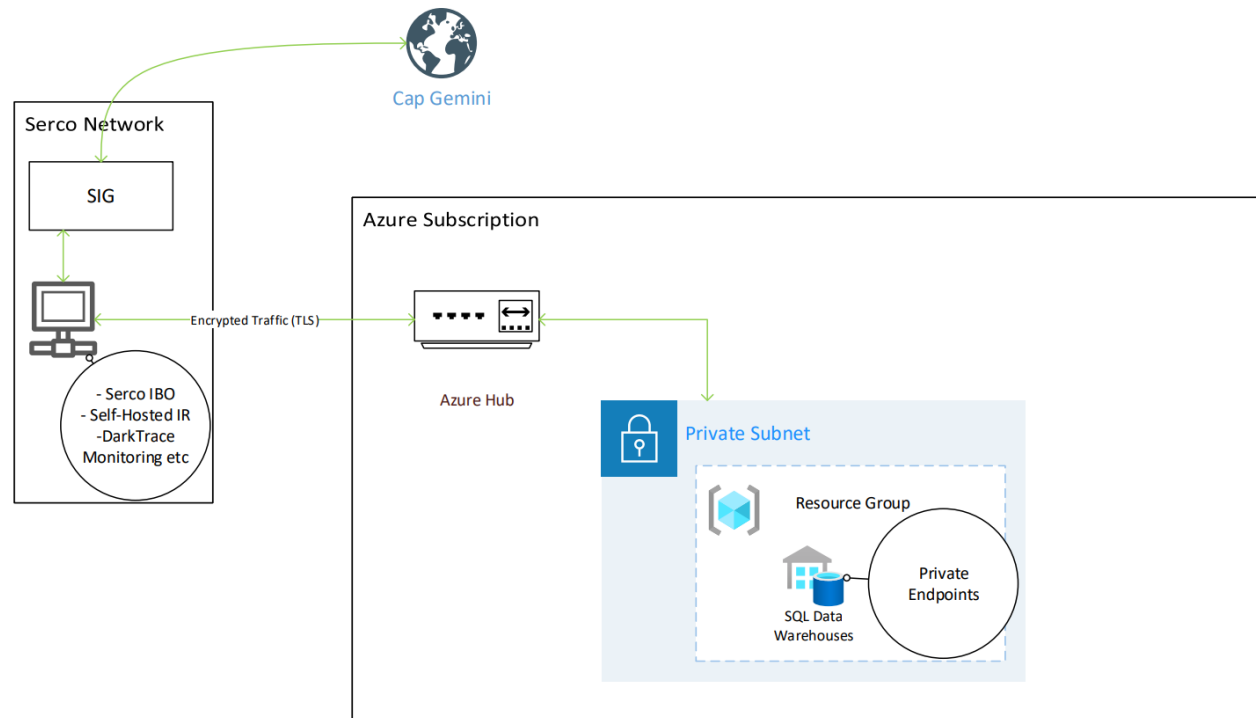
AZURE DATA SCIENCE ARCHITECTURE

- 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

The screenshot illustrates the process of connecting to a Parquet file in Power BI Desktop. The 'Get Data' pane on the left shows the 'Parquet' connector selected. The 'Parquet' dialog box in the center shows the file path 'C:\Users\sarahmed\Downloads\userdata1.parquet'. A preview window titled 'userdata1.parquet' displays a table of user data. The 'Visualizations' pane on the right shows a bar chart visualization.

registration_datetime	id	first_name	last_name	email	gender	ip_address	cc	country
3/02/2016 7:55:29 AM	1	Amanda	Jordan	ajordan0@com.com	Female	1.197.201.2	6759521864920116	Indonesia
3/02/2016 5:04:03 PM	2	Albert	Freeman	afreeman1@is.gd	Male	218.111.175.34		Canada
3/02/2016 1:09:31 AM	3	Evelyn	Morgan	emorgan2@altavista.org	Female	7.161.136.94	6767119071901597	Russia
3/02/2016 12:36:21 AM	4	Denise	Riley	driley3@gmpg.org	Female	140.35.109.83	3576031598965625	China
3/02/2016 5:09:31 AM	5	Carlos	Burns	cburns4@mitelban.gov.cn		169.113.235.40	5602256255204850	South Africa
3/02/2016 7:22:34 AM	6	Kathryn	White	kwhite5@google.com	Female	195.131.81.179	3583136326049310	Indonesia
3/02/2016 8:33:08 AM	7	Samuel	Holmes	sholmes6@fomeuws.com	Male	232.234.81.197	3582641366974690	Portugal
3/02/2016 6:47:06 AM	8	Harry	Howell	hhowell7@eepurl.com	Male	91.235.51.73		Bosnia and Her
3/02/2016 3:52:53 AM	9	Jose	Foster	jfoster8@yelp.com	Male	132.31.53.61		South Korea
3/02/2016 6:29:47 PM	10	Emily	Stewart	estewart9@opensource.org	Female	143.28.251.245	8574254110301671	Nigeria
3/02/2016 12:10:42 AM	11	Susan	Perkins	sperkinsa@patch.com	Female	180.85.0.62	8573823609854134	Russia
3/02/2016 6:04:34 PM	12	Alice	Berry	aberryb@wikipedia.org	Female	246.225.12.189	4917830851454417	China
3/02/2016 6:48:17 PM	13	Justin	Berry	jberryc@usatoday.com	Male	157.7.146.43	6331109912871813274	Zambia
3/02/2016 9:46:52 PM	14	Kathy	Reynolds	kreynoldsd@redcross.org	Female	81.254.172.13	5537178462965976	Bosnia and Her
3/02/2016 8:53:23 AM	15	Dorothy	Hudson	dHUDSONe@blogger.com	Female	8.59.7.0	354258882824170	Japan
3/02/2016 12:44:01 AM	16	Bruce	Willis	bwillisf@bluehost.com	Male	239.182.219.189	3573030625927601	Brazil
3/02/2016 12:57:45 AM	17	Emily	Andrews	eandrewsg@cornell.edu	Female	29.231.180.172	30271790537626	Russia
3/02/2016 4:44:24 PM	18	Stephen	Wallace	swallaceh@netvibes.com	Male	152.49.213.62	5433943468526428	Ukraine
3/02/2016 11:45:54 AM	19	Clarence	Lawson	clawsoni@kontakte.ru	Male	107.175.15.152	3544052814080964	Russia
3/02/2016 10:30:36 AM	20	Rebecca	Bell	rbellj@bandcamp.com	Female	172.215.104.127		China

- 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



INTRODUCTION

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

Serco Data Science POC: Discovery Report | 6 December 2022

MACHINE LEARNING TECHNIQUE RESEARCH

XGBoost / Gradient Boosting and the Search Algorithms

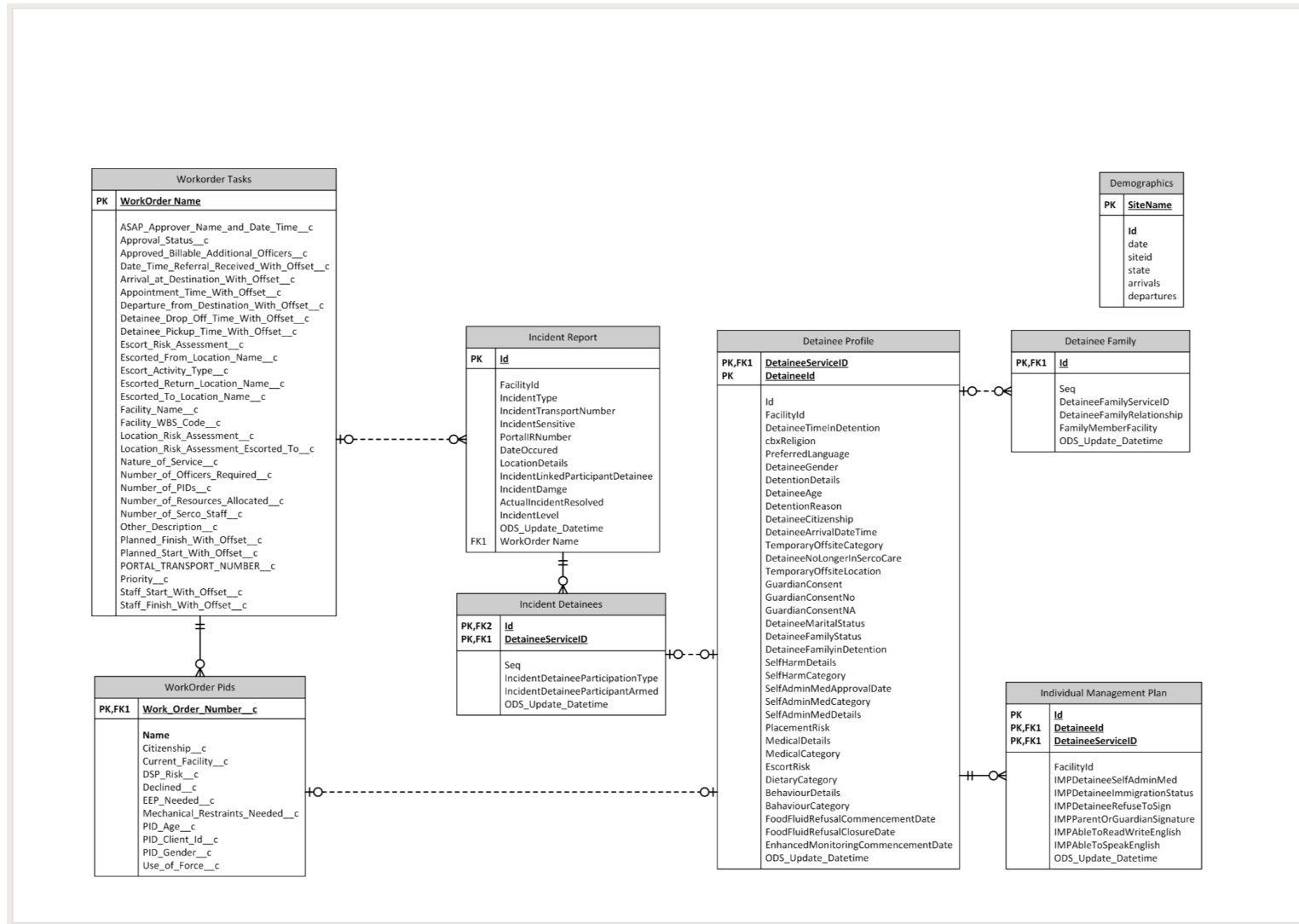
Tool	Description	Potential in predictive performance for risk assessment
XGBoost "Extreme Gradient Boosting"	<ul style="list-style-type: none">• XGBoost works by adding a classifier at a time, building upon the previous tree structure, whereas RF randomly and independently trains the classifier.• In other words, the boosted tree model is based on a sequential logic in which each new tree utilizes information gained by the previous trees.	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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.	<ul style="list-style-type: none">• 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

Company Confidential © Capgemini 2022. All rights reserved | 8



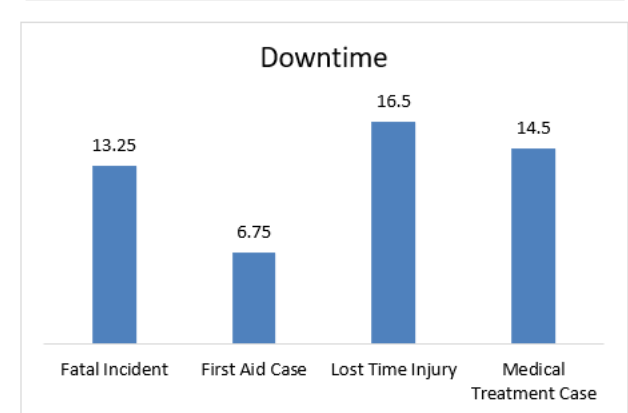
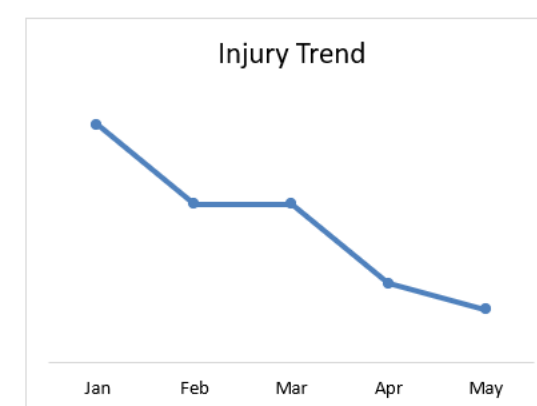
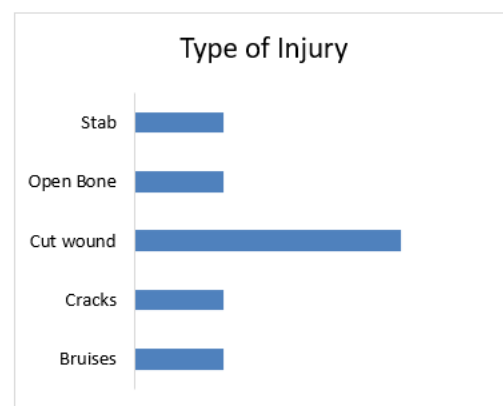
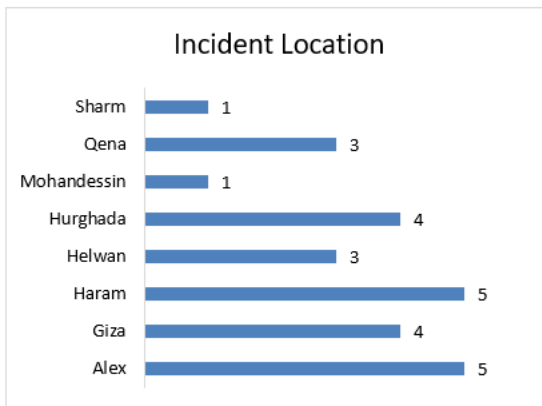
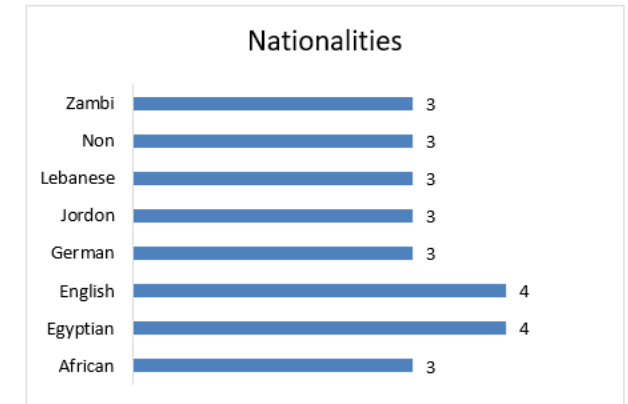
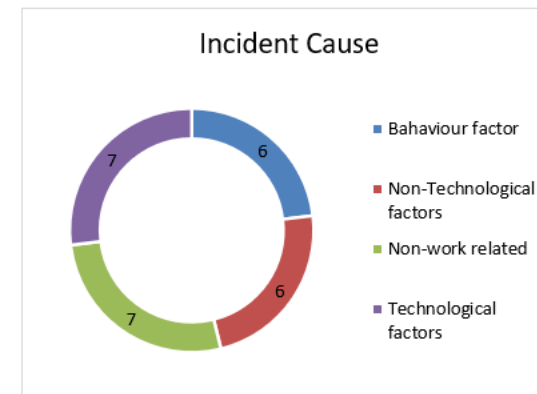
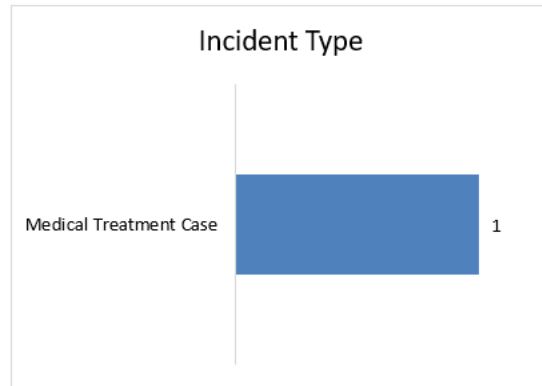
INPUT DATA MODEL



- 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

Microsoft Azure

Search resources, services, and docs (G+)

📧

🔍

🔔

⚙️

❓

🗨️

Ben.Moretti@serco-ap.c...

SERCO (SERCO.ONMICROSOFT.C...

Home > srappstabilcapgemini | Containers >

raw

Container

Search

«

📁 Upload

➕ Add Directory

🔄 Refresh

↻ Rename

🗑️ Delete

↔️ Change tier

🔑 Acquire lease

🔓 Break lease

Overview

Diagnose and solve problems

Access Control (IAM)

Settings

Shared access tokens

Manage ACL

Access policy

Properties

Metadata

Authentication method: Access key (Switch to Azure AD User Account)

Location: raw / Case Management

Search blobs by prefix (case-sensitive)

Show deleted objects

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

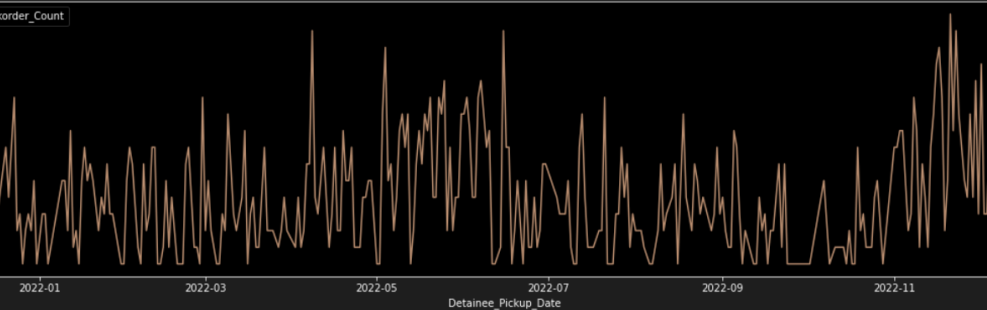


```
Cmd 23
```

```
# 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]:

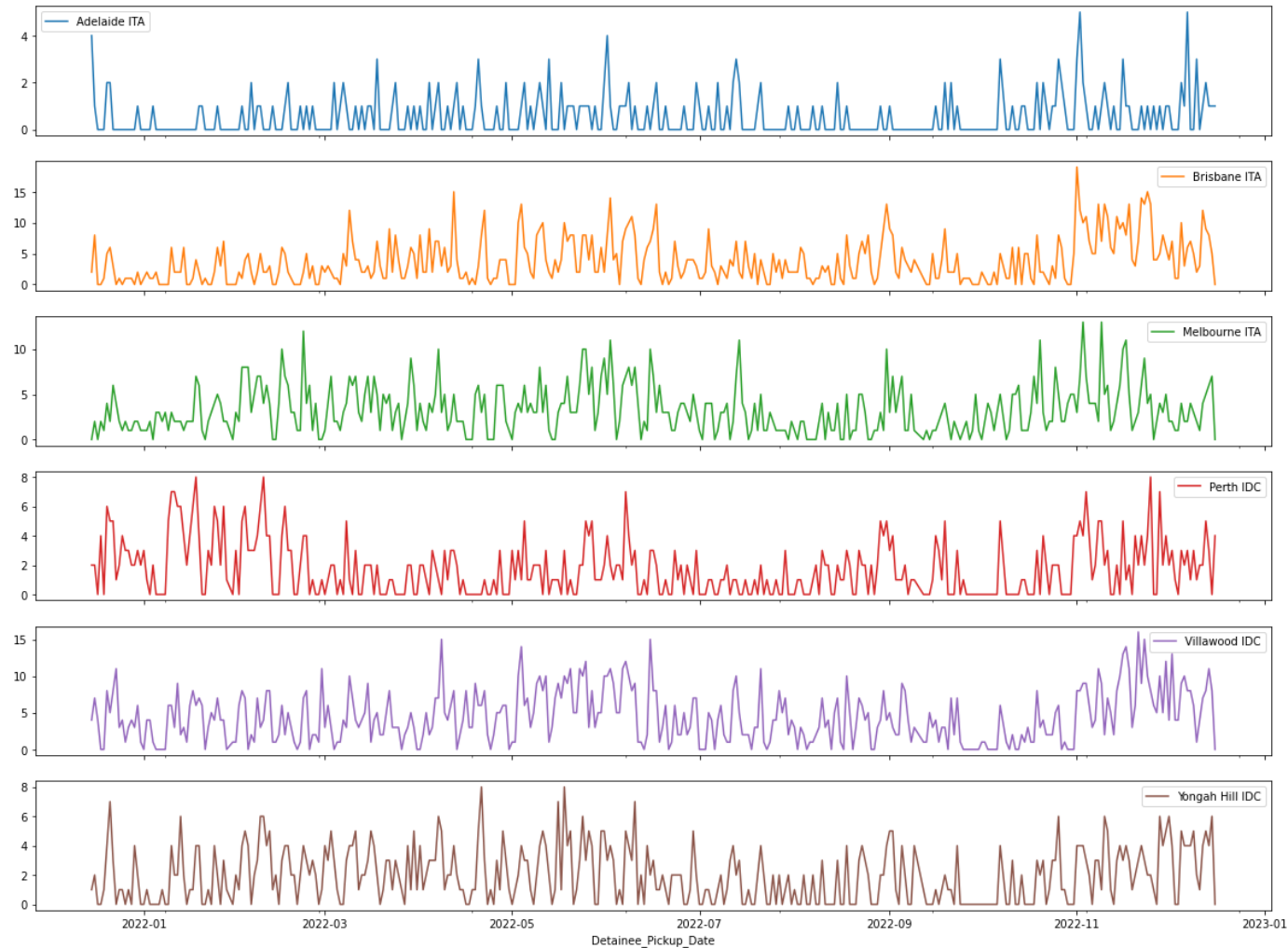


<AxesSubplot:title={'center':'Data Set: villawood_idc_workorders_sdf. Time Span: 1 year'}, xlabel='Detainee_Pickup_Date'>

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

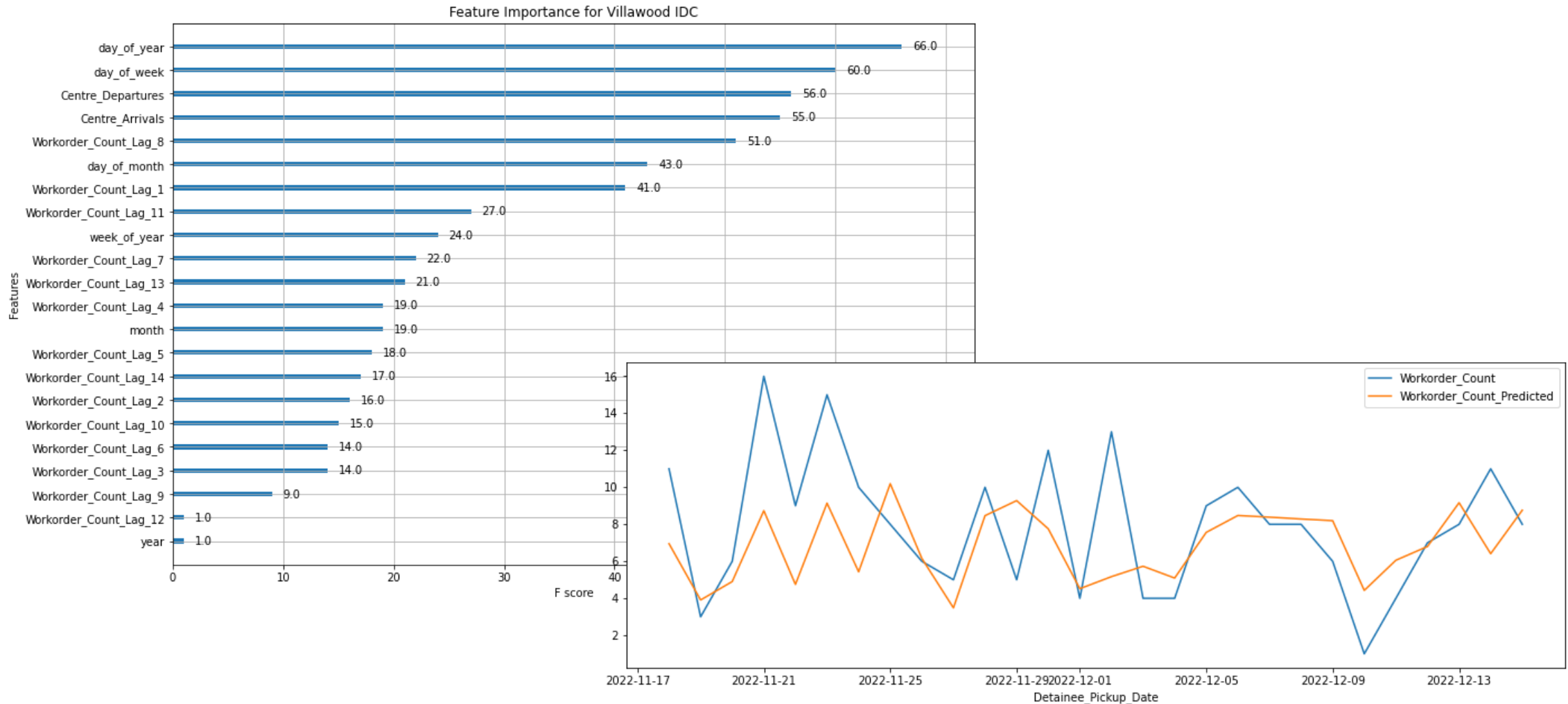


T & E EVENTS OVER TIME



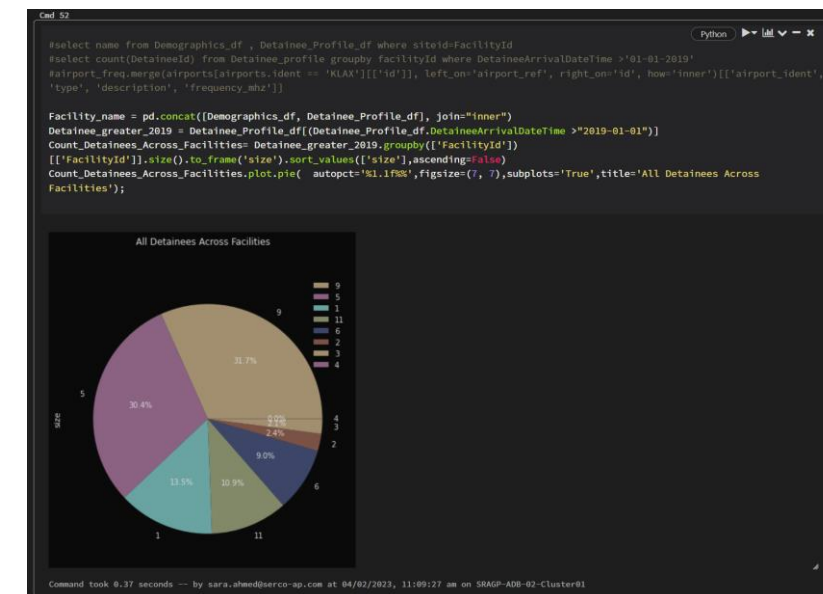
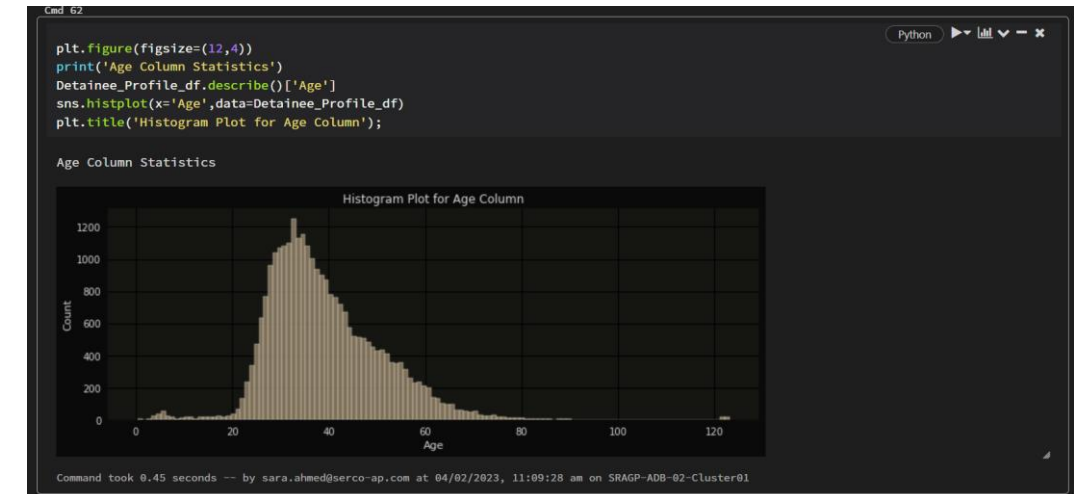
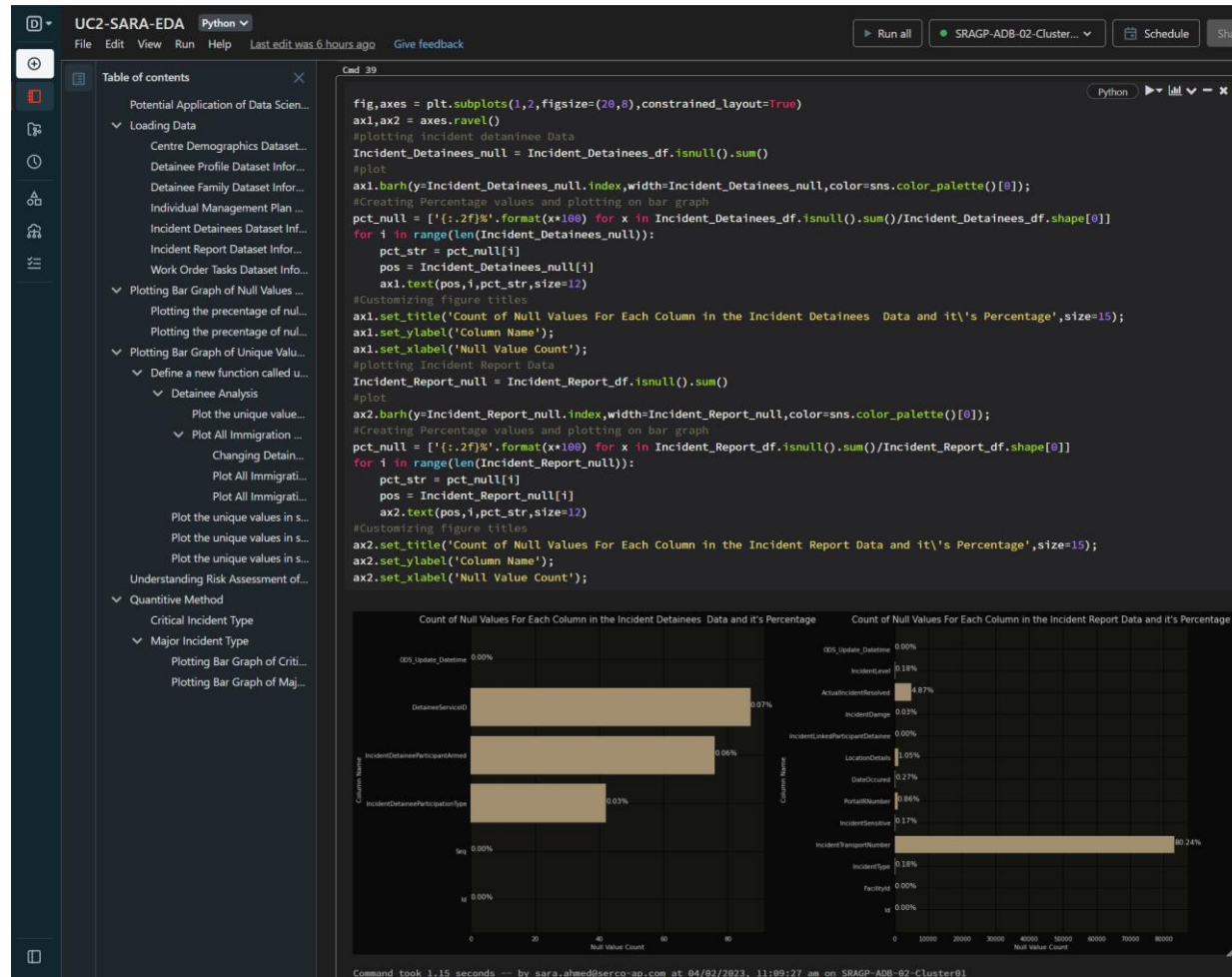


UC1 TRANSPORT AND ESCORT DEMAND MODEL





UC2 INCIDENT FORECASTING MODEL





This presentation contains information that may be privileged or confidential and is the property of the Capgemini Group.

Copyright © 2022 Capgemini. All rights reserved.