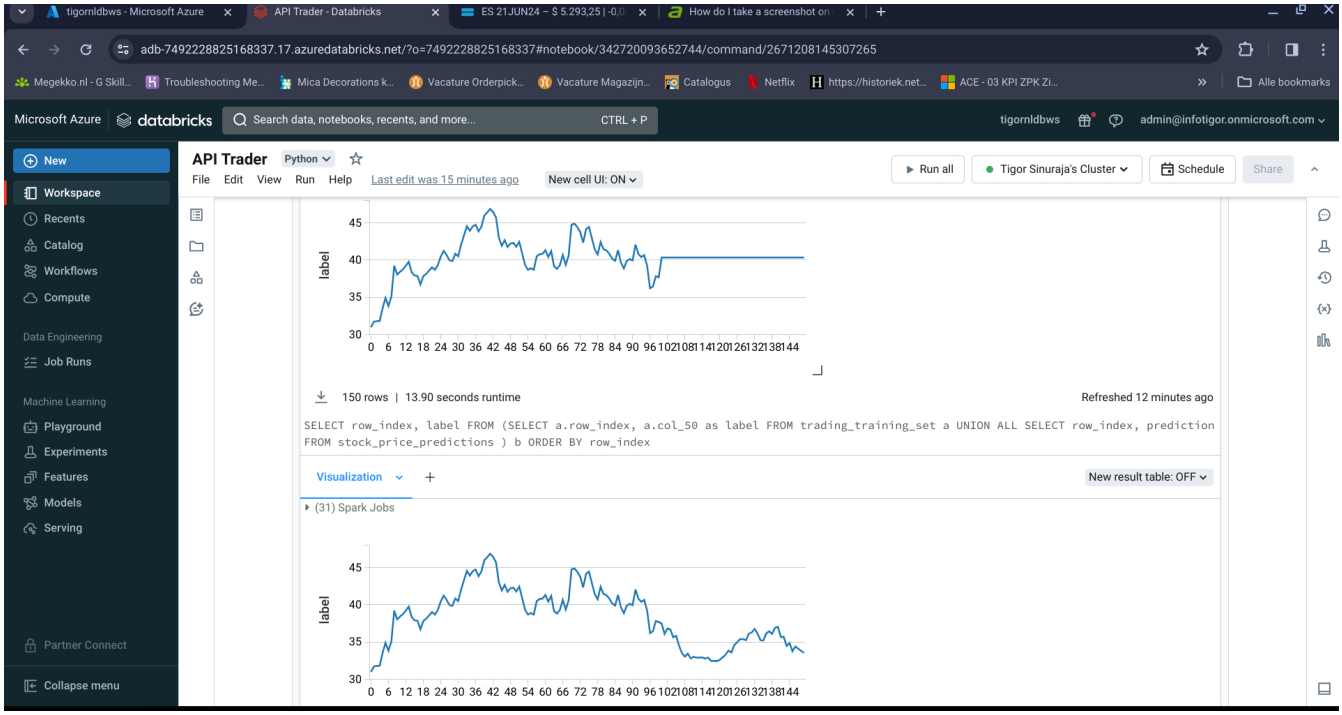
2024-03-23 Saturday hackaton: streaming data, Delta Tables, Machine Learning n Databricks
notebooks found from 2 years ago.

A couple of Saturdays were dedicated to get streaming system going using Structured Streaming APIs with FlightRadar24 data in Databricks. In that proces I found notebooks I created two years ago. Actually, I was surprised by things I could do :-). Using degiroapi get data from the DeGiro site, the stock price of Roblox and using pyspark and Random Forest model I predicted the stock price. And it still works! although I had to hack degiroapi a bit.

Oh yeah, and I bought Microsoft Support for my subscription. My first experience was actually great. I was immedately helped by 2 support people and afterwards I was interviewed by their manager and he asked if I was satisfied. Yes, I was!



2024-03-10 Databricks and DBT

Lately I have been looking at Databricks again. With my recent experiences with PySpark I want to use it in a modern scalable and easy to use framework. To my surprise Databaricks has SQL datawarehouse en integraion with DBT. Basically they want to make it easy who know SQL (which is basically everyone) to use Databricks which totally makes sense. The only thing which is funny about it is that you dont need PysPark at all, at least if you are a Data Engineer. OKay. But I want something else than SQL. Otherwise, it is just like Snowflake, which is great by the way.

But I want to play with PySPark :-). Somehow, being busy with no Low Code developement in Visual Studio Code and Azure Data Studio makes me feel like developer again. And thats a good feeling. Whith my recent interactions with ChatGPT it also is a lot of fun. I guess you have to be Data Scientist to use Python and PySpark. Thats where Databricks is superior to Snowflake and Synapse. So be it.

I am going to go ahead and devote my Saturday Hackaton (probably more than one) to making a system which takes a streaming data and predicts something.

To be continued...

2023-04-06 What if ?

(unsolicited advice of an old man to those who decide)

What if I was a Head BI and I have to choose an architecture, BI-platform and tooling ? Green field, so to speak. What would be the winning combination ?

Having worked with a number of Dutch companies. Having hands-on experience with software and technologies like SAS, Microsoft Azure, Snowfake, Oracle, Hadoop, Spark, Informatica etc. Having done that on-prem and in the Cloud. Having build datawarehouses on relational databases and Data Lakes. Having modeled with Data Vault and without it.

I can say:

it's not an easy choice to make!

Goal

If I have to sum up what I (as a hypothetical Head of BI) would have wanted to achieve:

- Maximum output: you want to be able to have tangible results as quick as it is possible
- Minimum cost: that includes time (=money) to learn new skills, actually develop and deliver products
- Maintainability: you can replace people easily and you don't have to keep them because they know how one app works
- Continuity: make sure whatever you choose and buy, you don"t have to reconsider within one year

Challenges

Why is that so difficult?

- There is no Silver Bullet. There is no single solution for all. But there is probably objectively speaking one optimal choice for your specific situation.
- Datawarehouse, Data Lake or something else? That's an important one. Will talk about it in a separate chapter.
- The number of tools is growing. New vendors coming up continuously, little real experience with those tools is available, but they promise wonders
- Tools are evaluating. What seemed like an optimal choice yesterday can be not so obvious tomorrow
- Information about those tools is sometimes contradictory and depends on where it comes from. Benchmarks are often biased

Basic principles

Of course, if you have to make a decision of this magnitude, it has to be based on a clear vision and principles and if possible, real life facts and experience

- Hire and keep a good architect :-) Errors in design are very expensive to correct if the system is already build
- Keep it simple when it comes to structuring and modeling your Datawarehouse or Data Lake. Do you understand what a Data Vault is, what problem it solves en do you have that problems

- Keep it simple when organizing development process. Being developer myself for 20+ years I love making things. I love developing, low code or not, making code generators that make code. It is all good until someone else has to take over. For them it's often much less fun. Besides, your company suddenly confronted with the fact that it created critical software which it depends on and it has to maintain and which is not the primary business at all. So avoid it at all cost!
- Make or buy: definitely, if its available, BUY! see previous point.
- No one has canceled basic principles if you have to develop:
- Enforce strict rules and guidelines for development. It does not matter if you do coding or use low code. Otherwise you will get well known spaghetti
- Maintain reusable code in one place. Do not copy
- SQL is the only skill here to stay. Organize your development process around it and use it as much as possible. For instance use SparkSQL
- Open Source is only good when you pay for it, and you will, one way or another.
- You don't want to depend on a group of technical people who love what they are doing (like myself) but who are the only ones who know things. So go for a managed cloud service whenever possible. Yes, Cloud is not cheap, but in the end, it is cheaper! Yes, go Cloud.

to be continued...

2023-04-06 Unfortunately, Betfair is no longer available in NL putting my MMA -betting project on hold indefinitely. Instead, I am working on an automatic prediction and ordering system based on Degiro. This is done by endless number of people before. What I am aiming for is to use Azure managed cloud services to full extend. I wonder what it would mean in terms of development effort, maintainability and cost,

- Getting historical training data - Azure Function (Python). The main argument to use this technology besides the obvious leaning effect was of course the serverless nature of Azure functions. As the whole point was to leverage the power of Cloud managed services this is a logical choice. My observations so far:

- If you use python, you cant edit in the browser. Developing in Visual Studio Code locally is an excellent alternative though
- To make things work after deployment like importing modules was quite painful even after reading numerous docs and blogs, I would like it to be much easier to handle, It all worked in the end, but it costed to much time. I want to get a result and I don't care what path I have to set or config I have to fill. It all should be hidden and just work!
- in the case of degiroapi I had to alter request method (otherwise, the site thinks its a bot and refuses connection). Also here it felt like time wasted if it wasn't for the learning effect of it.

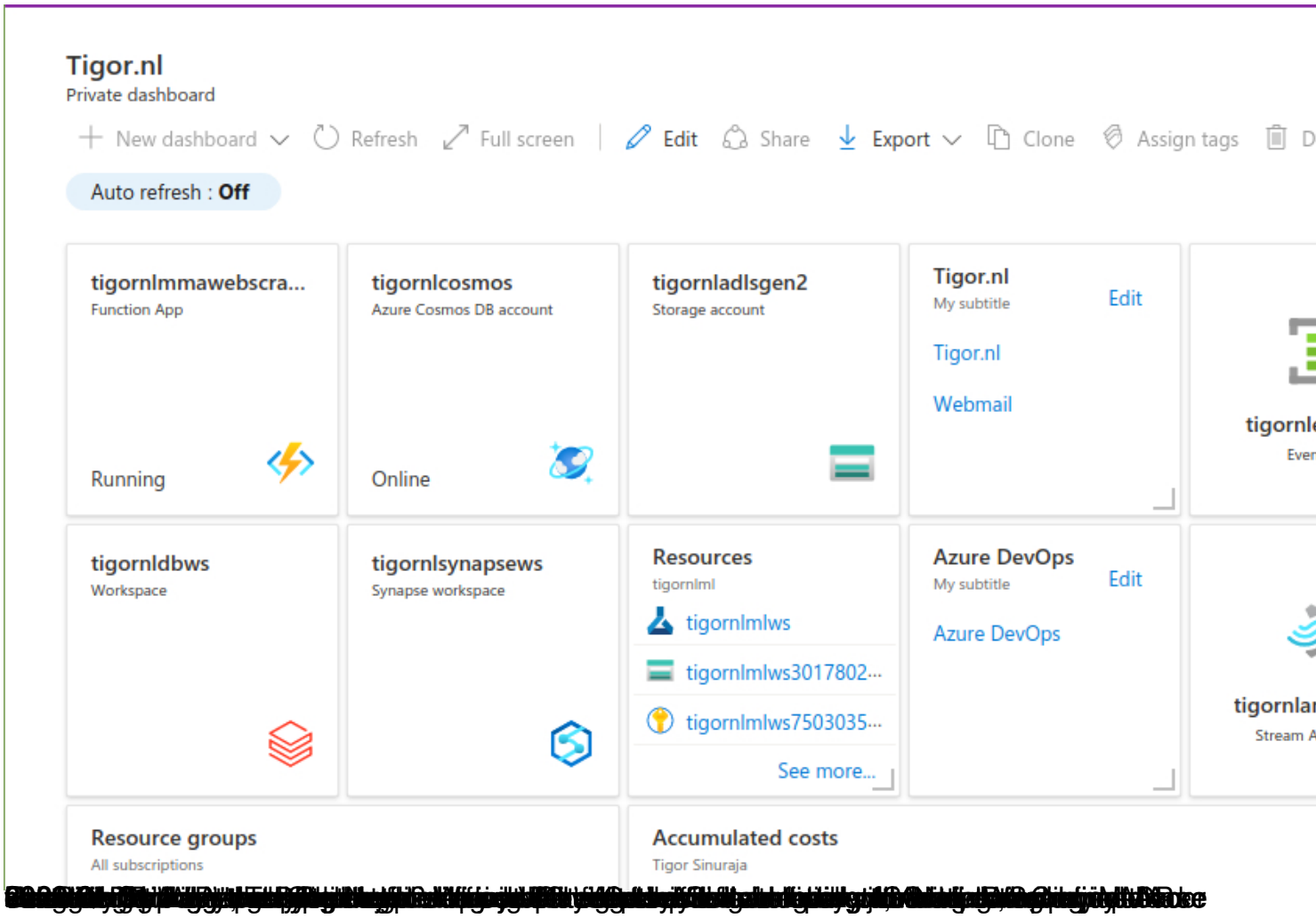
- Row data layer - we store collected data in Azure Cosmos DB, JSON files as documents, later I will probably transition to pure Data Lake solution.
- Permanent storage, historization for tranformed data will go to Delta Lake - exciting to use time travel and ACID features
- Data preparation - Synapse, pyspark notebooks, in my opinion, this can be used as alternative to Databricks (for the purpose in mind) and Azure Data Factory. The latter functionality is integrated in Synapse. Dedicated pools I wont use because of the cost. Mainly, its the combination of files in Data Lake as storage and Spark pools for compute.

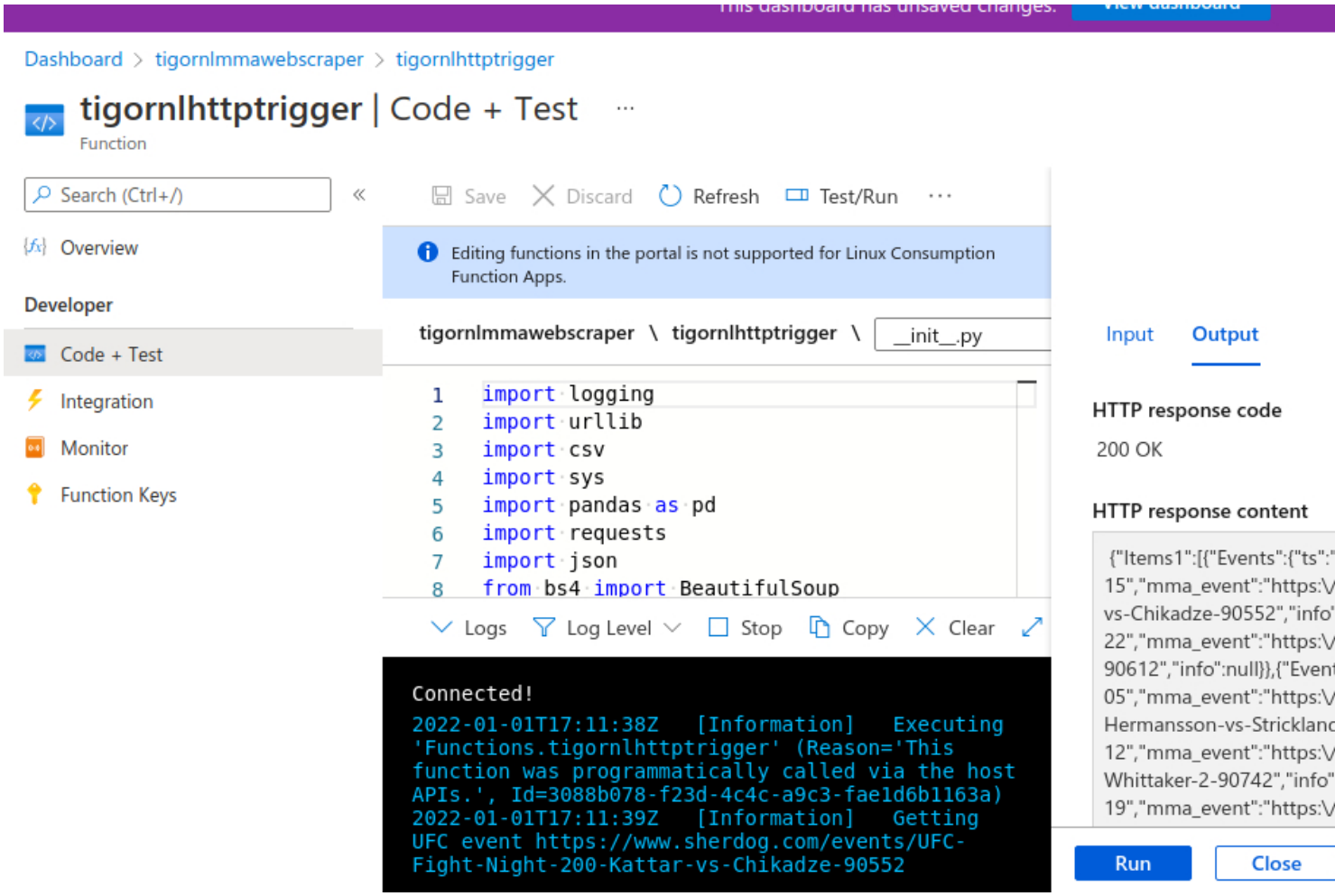
- Training a model and generate inference model - Azure Automated ML as low code as possible(later may be Databricks and Azure ML SDK)

- Getting instances to predict - Azure Function (Python)
- Ad hoc analysis - Azure Synapse Serverless pool
- Visualization - PowerBI, integrated with Synapse
- Versioning and deployment - Azure Devops

2022-01-02 An idea how the Cloud version of my application is going to be.

- WebScraping of training data - Azure Function (Python)
- Store collected data - Azure Cosmos DB
- Data preparation - Databricks and Azure Data Factory
- Training a model and generate inference model - Azure Automated ML as low code as possible(later may be Databricks and Azure ML SDK)
- Getting instances to predict - Azure Function (Python)
- Ad hoc analysis - Azure Synapse Serverless pool
- Visualization - PowerBI
- Versioning and deployment - Azure Devops


Tigor.nl
Private dashboard
New dashboard
Refresh
Full screen
Edit
Share
Export
Clone
Assign tags
Auto refresh : Off
tigornlmmawebscra...
Function App
Running
tigornlcosmos
Azure Cosmos DB account
Online
tigornladlsgen2
Storage account
Tigor.nl
My subtitle
Edit
Tigor.nl
Webmail
tigornldbws
Workspace
tigornlsynapsews
Synapse workspace
Resources
tigornlml
tigornlmlws
tigornlmlws3017802...
tigornlmlws7503035...
See more...
Azure DevOps
My subtitle
Edit
Azure DevOps
Resource groups
All subscriptions
Accumulated costs
Tigor Sinuraja


This dashboard has unsaved changes.
Dashboard > tigornlmmawebscraper > tigornlhttptrigger
tigornlhttptrigger | Code + Test
Function
Search (Ctrl+/)
Overview
Developer
Code + Test
Integration
Monitor
Function Keys
Save
Discard
Refresh
Test/Run
Editing functions in the portal is not supported for Linux Consumption Function Apps.
tigornlmmawebscraper \ tigornlhttptrigger \ __init__.py
1 import logging
2 import urllib
3 import csv
4 import sys
5 import pandas as pd
6 import requests
7 import json
8 from bs4 import BeautifulSoup
Logs
Log Level
Stop
Copy
Clear
Connected!
2022-01-01T17:11:38Z [Information] Executing 'Functions.tigornlhttptrigger' (Reason='This function was programmatically called via the host APIs.', Id=3088b078-f23d-4c4c-a9c3-fae1d6b1163a)
2022-01-01T17:11:39Z [Information] Getting UFC event https://www.sherdog.com/events/UFC-Fight-Night-200-Kattar-vs-Chikadze-90552
Input
Output
HTTP response code
200 OK
HTTP response content
{"Items1":[{"Events":{"ts":"
15","mma_event":"https:\/
vs-Chikadze-90552","info"
22","mma_event":"https:\/
90612","info":null}},{"Event
05","mma_event":"https:\/
Hermansson-vs-Strickland
12","mma_event":"https:\/
Whittaker-2-90742","info"
19","mma_event":"https:\/
Run
Close