Using Azure Machine Learning in Stream Analytics

U2U

Nico Jacobs, U2U training
@SQLWaldorf

Dr. Nico Jacobs

Trainer at U2U

/sqlwaldorf

@sqlwaldorf

University of Leuven, Belgium

Studied computer science
Teaching assistant for 7 years
Ph.D. in machine learning

Trainer at U2U

Trainer at U2U since 2004:

SQL Server 2000 and later
SSIS, SSAS, SSRS, …
Power BI
Azure Big Data

Speaker at IT events:

SQLBits, PASS Summit, TechDays, Techorama,
SQL Saturdays, DataMinds Connect, ESPC, …

Agenda


Azure
Stream Analytics


Azure
Machine Learning


Combining both


Conclusions

PASS

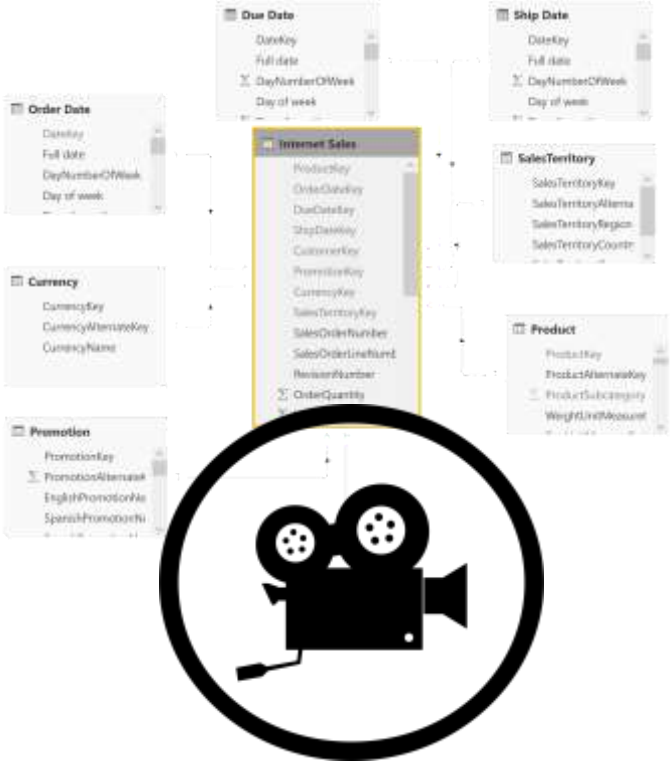
Azure Stream Analytics

PASS

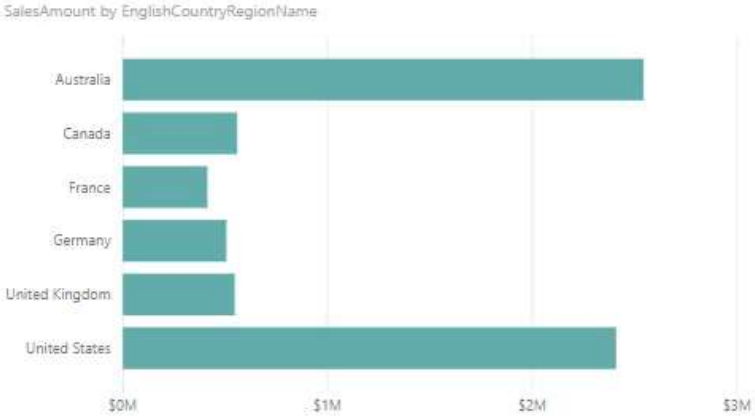
‘Traditional’ analytics

SQL, DAX, MDX query

Stream analytics

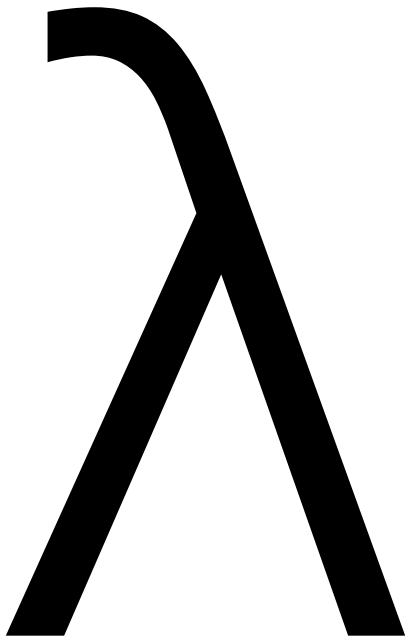


ASA SQL query



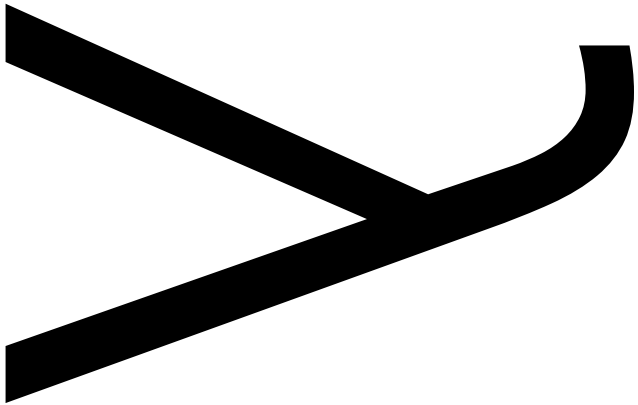
PASS

Lambda architecture



Lambda architecture

Realtime data
Scoring

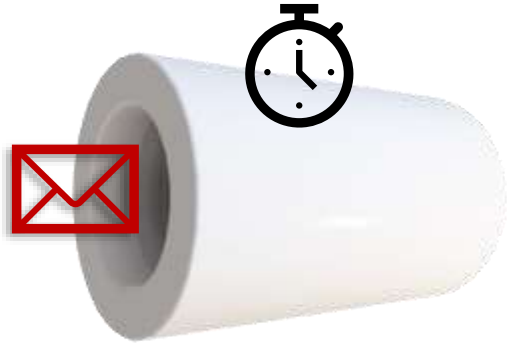


Historical data
Training

Inputs

Fact tables → Stream input

- Event Hubs
- IoT Hubs
- Blob storage



Support for compressed data

For ultimate performance take partitions into account

Inputs

Dimension table → Reference input

Blob storage or Azure SQL Database

Either static or slowly changing data

Files in Year/Month/Day/… structure or query on temporal table

Limited support for reference data:

- 50 Mb with 1 streaming unit, up to 300 Mb with 6 su

Outputs

After processing the input, one or more output streams can be sent to a wide variety of outputs:

Queues: Event Hub, Service Bus topics & queues

Database: SQL database, Cosmos DB

Storage: Blob, Data Lake, Table

Processing: Azure function

Visualization: Power BI

Query: connect input & output

ASA maps input(s) to output(s) using a SQL query

- Rapid development because little and easy coding
- Subset of T-SQL language
- Additional functions and keywords to deal with temporal aspect

Time is an illusion…

Exactly which time is used?

- EnqueuedUtcTime for Event & IoT Hub
- Last modified date for Blob

To change this use TIMESTAMP BY in FROM clause

How to aggregate an infinite stream?

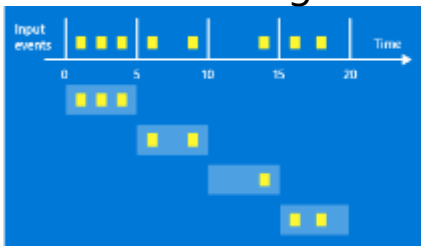
Input is a potential infinite stream

Aggregated values only possible when limiting the timespan over which they are computed

4 windowing functions in ASA

4 Windowing functions

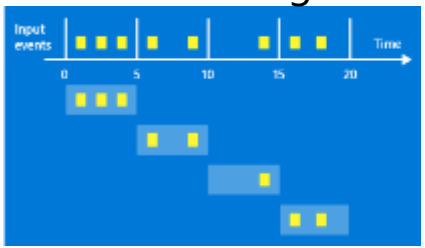
Tumbling



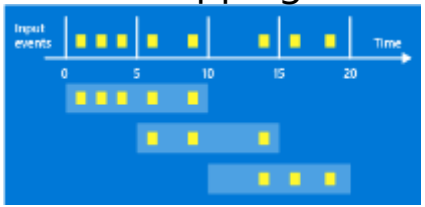

```
SELECT sum(Toll), TollId, System.Timestamp as ts
FROM bootin
TIMESTAMP BY EntryTime
GROUP BY tumblingwindow(minute, 5), TollId;
```

4 Windowing functions

Tumbling



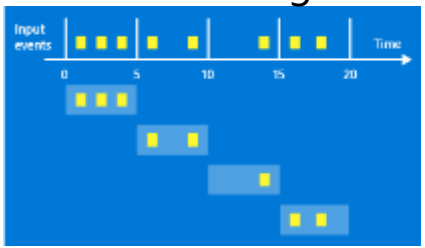
Hopping



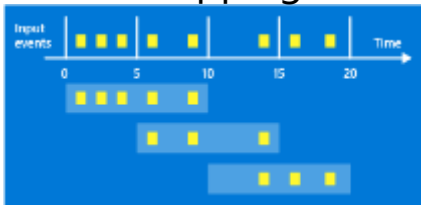

```
SELECT sum(Toll), TollId, System.Timestamp as ts
FROM bootin
TIMESTAMP BY EntryTime
GROUP BY hoppingwindow(minute, 10, 5), TollId;
```

4 Windowing functions

Tumbling



Hopping



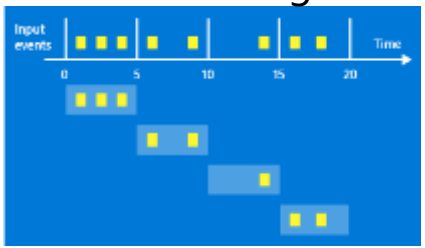
Sliding


```
SELECT sum(Toll), TollId, System.Timestamp as ts
FROM bootin
TIMESTAMP BY EntryTime
GROUP BY slidingwindow(minute, 5), TollId;
```

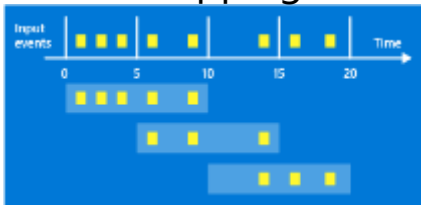
PASS

4 Windowing functions

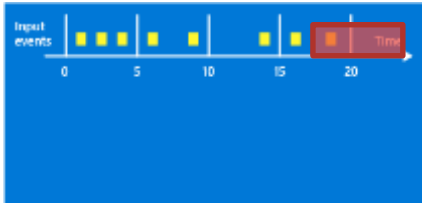
Tumbling



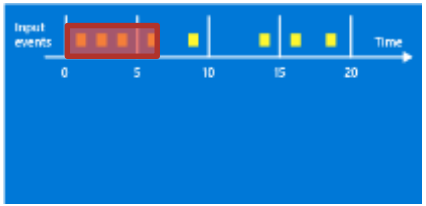
Hopping



Sliding



Session

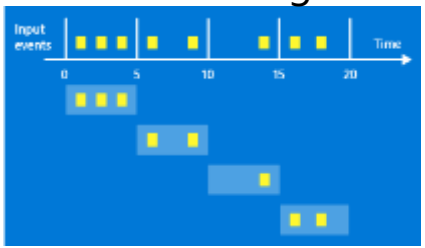



```
SELECT sum(Toll), TollId, System.Timestamp as ts
FROM bootin
TIMESTAMP BY EntryTime
GROUP BY sessionwindow(minute, 5, 60), TollId;
```

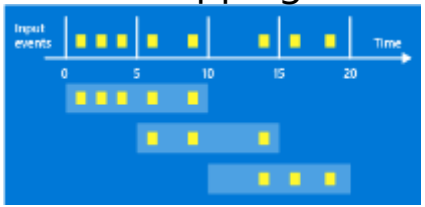
PASS

4 Windowing functions

Tumbling

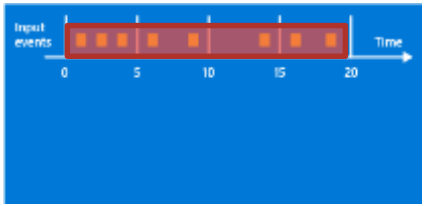


Hopping



Sliding

Session




```
SELECT sum(Toll), TollId, System.Timestamp as ts
FROM bootin
TIMESTAMP BY EntryTime
GROUP BY sessionwindow(minute, 5, 60), TollId;
```

PASS

Stream Analytics Query

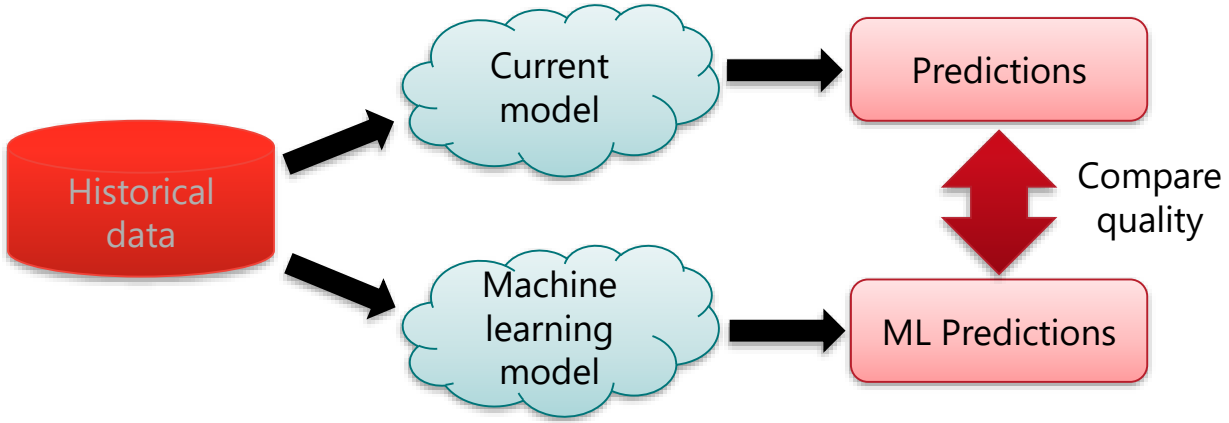
Demo

Azure Machine Learning

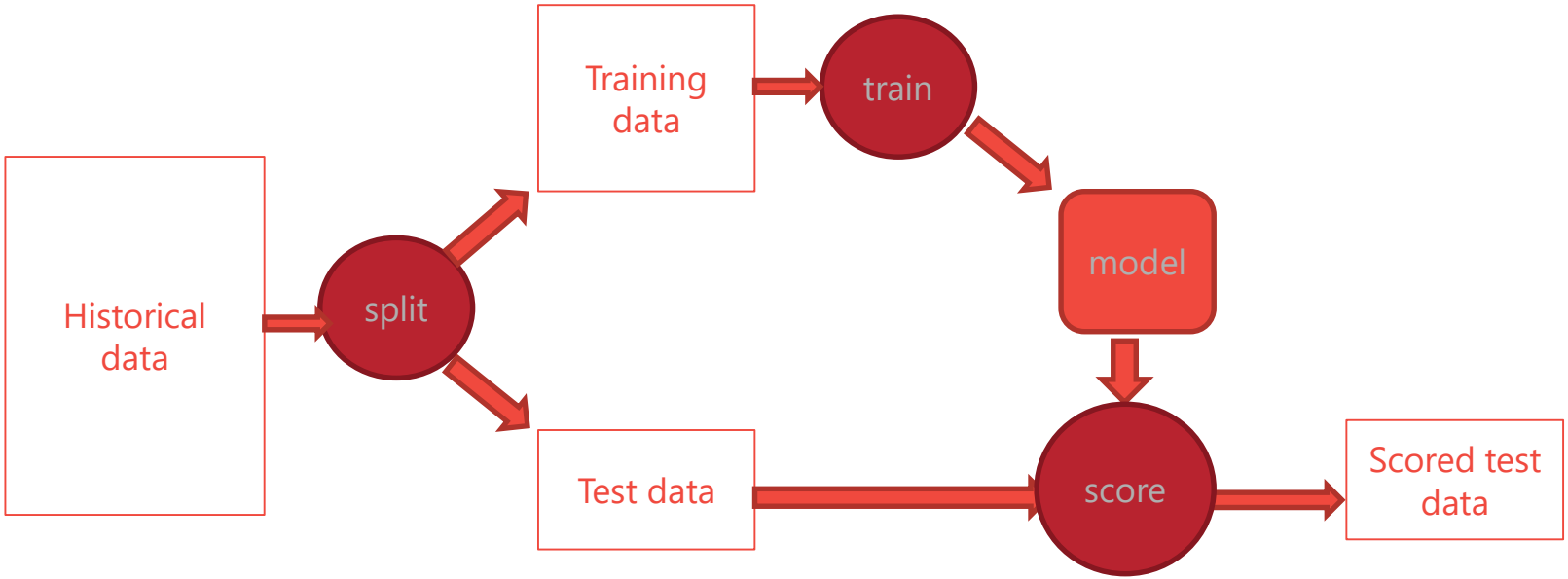
PASS

Why Machine Learning?

Verify if your current 'educated guessing' (a.k.a. model) can be improved by machine generated model



Training a model



The Azure Machine Learning twin

Two services:

- Azure Machine Learning Studio Classic
- Azure Machine Learning Services

Azure Machine Learning Studio Classic

Easy to use and deploy

Authoring via graphical designer

Focused on shallow learning

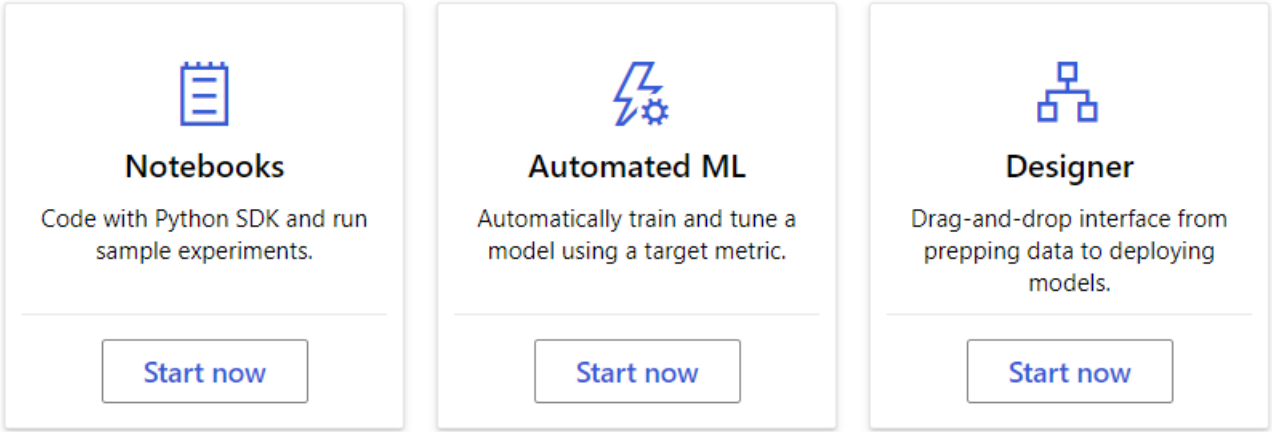
Stable but no further development

Azure Machine Learning Services

Recently released

Still a lot of innovation

Supports both shallow and deep learning



Azure Machine Learning

Demo

PASS

Combining both

PASS

Why combine both?

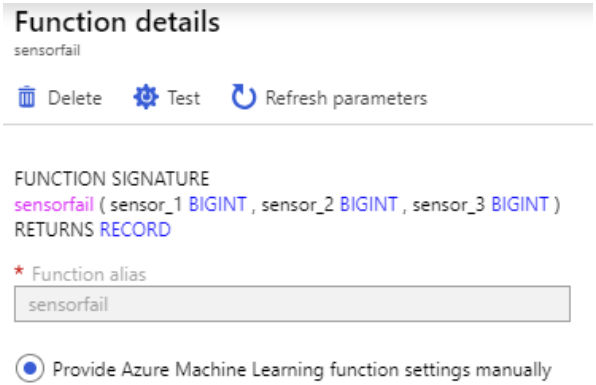
Often machine learning is better or cheaper in identifying patterns in data than humans

Same holds for keeping the patterns in sync with the constant changes in the world

How to combine them?

Azure Stream Analytics allows for functions to be called in the streaming query

- Custom JavaScript code
- Calls to Azure Machine Learning web services



PASS

Preparing data for Machine Learning

Different scenarios:

Predicting non-stream value based on current state

Predicting non-stream value based on time window

Predicting future stream value based on time window

Non-stream values based on current state

Example: sentiment score on text events

Requires manual labeling of training set

Every event can call web service: potential performance issue

Stream values based on time window

Example: predict future sensor reading based on previous readings

Historical scored data can easily be collected by Stream Analytics

Which values to use from which time window?

Non-stream values based on time window

Example: predict future device failure based on previous readings

Combines both challenges: manual labeling training data and decide which values to use from time window

Combining both

Demo

Performance

Azure Stream Analytics in general can improve performance in 3 ways:

- More streaming units (to some degree)
- Partitioning
- Common table expressions

Performance

When using Azure Machine Learning there are some extra concerns:

- Performance of the webservice
- Number of parallel calls to the ML webservice
- Batch size

Conclusions

PASS

Azure Stream Analytics

Easy way to analyze stream of events

Similar to traditional analytics

- Star schema as source
- Write a SQL query for analysis
- Use results in application

Azure Machine Learning

Machine learning helps taming the unknown

Azure Machine Learning is very intuitive to use

Deployment via the generated web service is easy

ASA & AML: better together

Stream Analytics benefits from a Machine Learning model when the processing is difficult to handcraft

Choose an appropriate data representation for model training

Thank You

@SQLWaldorf

Nico@u2u.be