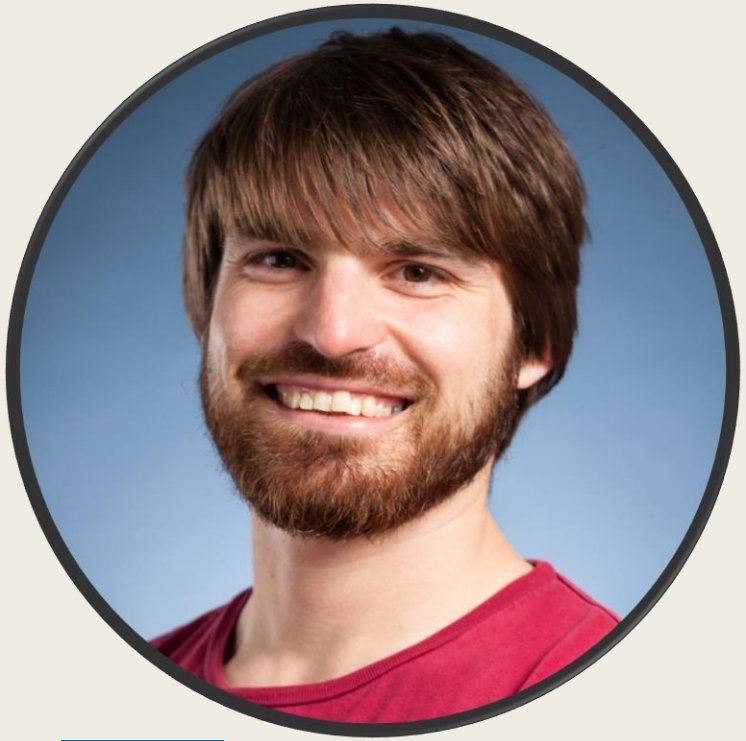
BUILDING AN ANALYTICS PLATFORM IN AZURE

Gerhard Brueckl

paiqo.com



Microsoft Most Valuable Professional

SSAS MAESTRO by Microsoft

@gbrueckl

blog.gbrueckl.at

gerhard@gbrueckl.at

https://github.com/gbrueckl

DatabricksPS

Databricks VSCode

paiqo

www.paiqo.com

PASS SQL saturday VIENNA

PASS AUSTRIA

Analytics Platform in Azure

- What is an Analytics Platform?
- Why build it on Azure?
- Main Components
- Scenarios

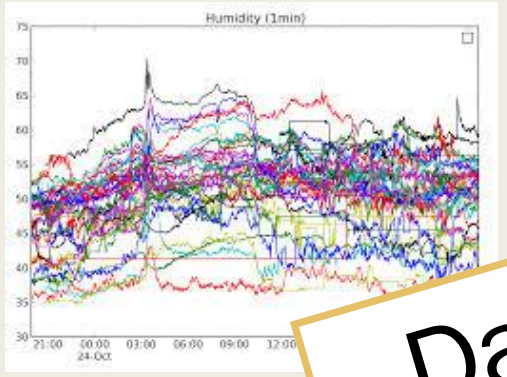
What is an Analytics Platform?

- Data Platform that integrates data from various sources for analytical purposes.
  - *SQL (ERP, CRM, …)*
  - *Big Data (logs, images, …)*
  - *Streaming / Real-Time (IoT, …)*
  - …
- Used by Data Scientists and Data Analysts
- Built by Data Engineers

What is an Analytics Platform?
History

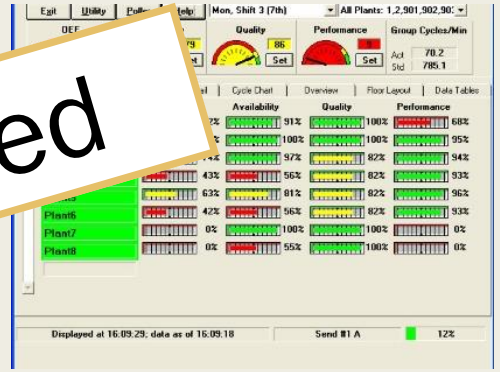
DWH

Regular Reports



Sensor/Log

Data was not stored



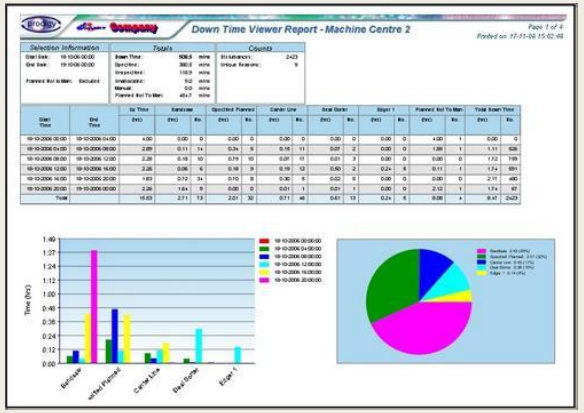
Live Reports

What is an Analytics Platform?
Big Data

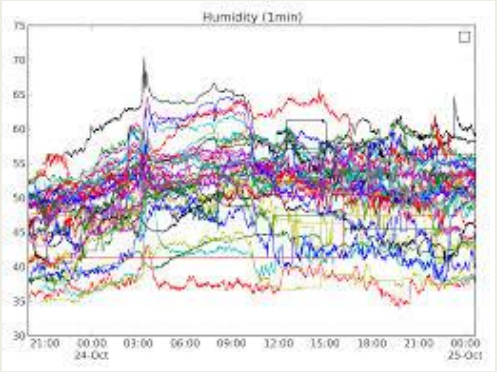
DWH

Regular Reports

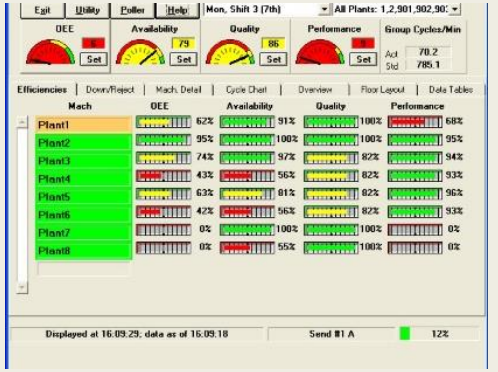
Archive



Operational Reports
e.g. for Maintenance



Sensor/Log Data

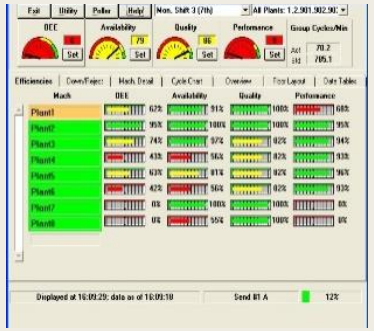


Live Reports

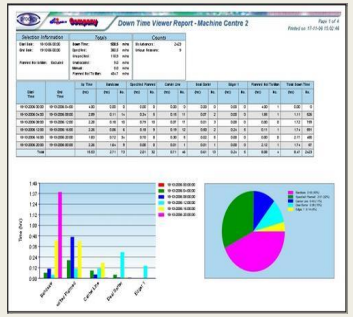
What is an Analytics Platform?
Advanced Analytics

DWH

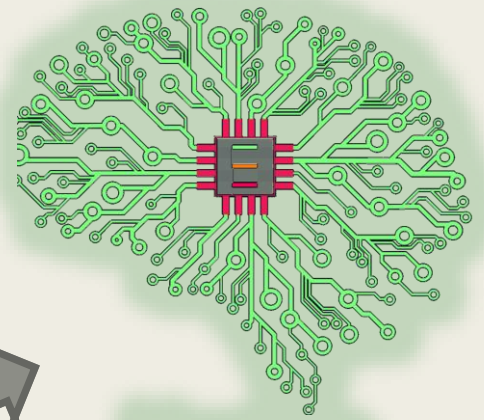
Regular Reports



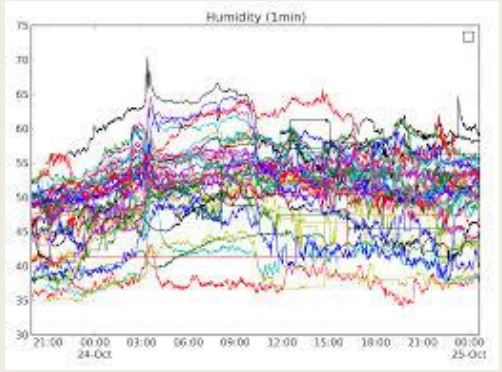
Live Reports



Operational Reports



Machine Learning



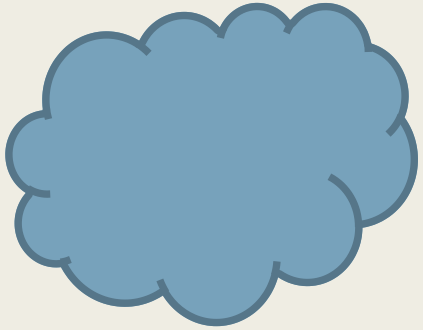
Sensor/Log Data

Data Lake

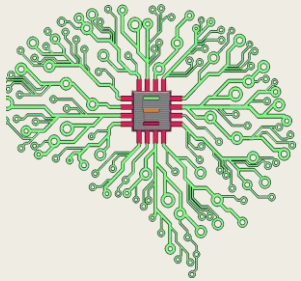
What is an Analytics Platform?
Cloud

DWH

Reporting



Data Lake



Machine Learning

Need for flexible, cost-efficient and high performant infrastructure

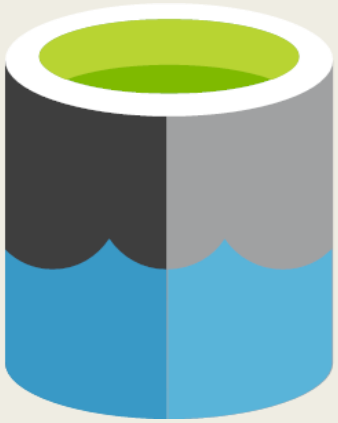
- use-cases
- D
- Different kind of w

Why build it on Azure?

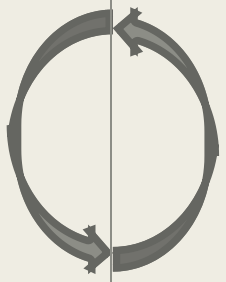
- "Grow at Scale" – no "Big Bang"
  - *Separation of Storage and Compute !!!*
- Could also be any other cloud provider
- BUT Azure has
  - *Best PaaS offerings*
  - *Best User Management*
  - *Best Overall Integration of Tools/Services*

Main Components

Data Management

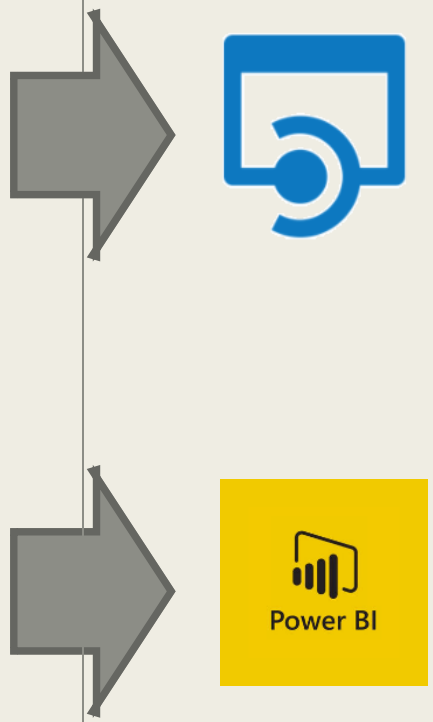


Storage



Processing

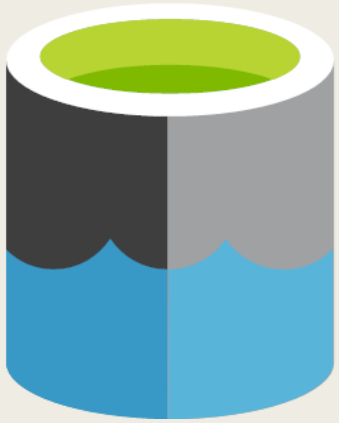
Serving



Consuming

Main Components - Storage

- Elastic / Flexible / Scalable
  - *Start small – grow big*
- Manageable
  - *Folders*
  - *Security Concept*
- (HDFS support)



Main Components - Processing

- Elastic / Flexible / Scalable
  - *On-demand clusters*
  - *Large Cluster-sizes (100+ nodes)*
  - *Large Nodes (256GB, 64 cores)*
  - *Structured & Unstructured data*

- Performance
  - *Distributed Processing*
  - *Fast Data Access*

Main Components - Processing

- Supported Engines / Languages
  - *Spark*
    - Python
    - Scala
    - R
    - C#
  - *SQL*

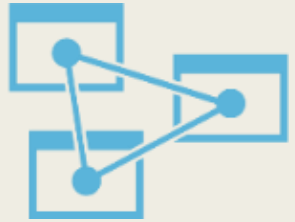
- Integrated Security
  - *Cluster Management*
  - *Data Access*

Main Components - Serving

- Hosting of ML models
- Kubernetes / Docker
- API Service
- Batch processing

Main Components - Serving

- Query-Interfaces
  - *SQL*
- OLAP-Interfaces
  - *Analysis Services*
  - *Power BI*
- File-Access



Main Components – Data Management

- Orchestration & Scheduling
  - *Data Pipelines*
  - *Connectivity / Extensibility*

- Meta Data Store
  - *API*
  - *Connectivity / Extensibility*

Getting Started - Storage

Can also be a combination!

Choose your Storage Service:

Data Lake Store Gen2

- Security
- Folders
- HDFS

Blob Storage

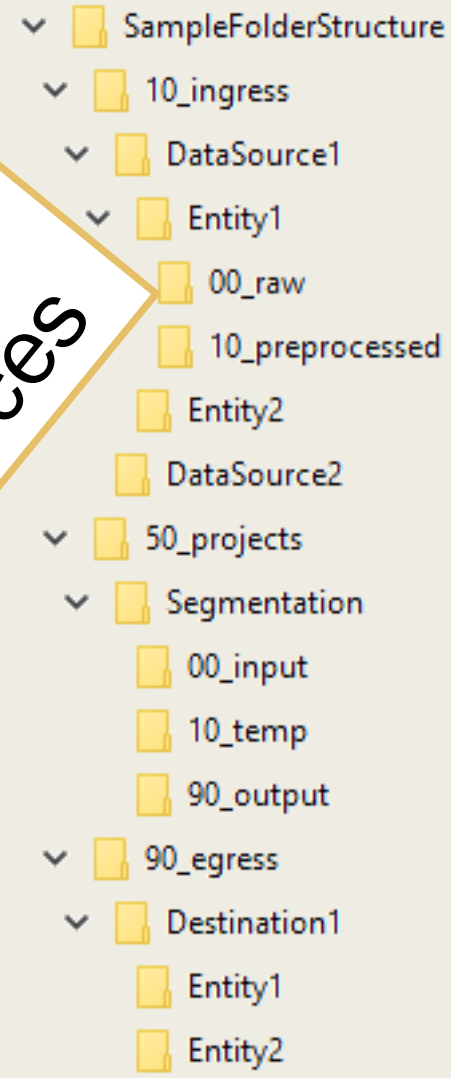
- Redundant
- Cheap
- Availability (Region)

Getting Started - Storage

Structure your Storage!

- Dedicated Areas for
  - *Input/raw data*
  - *Preprocessed/cleaned data*
  - *Working copies*
  - *Output data*
- Further partition by time (…/yyyy/MM/dd/…) !

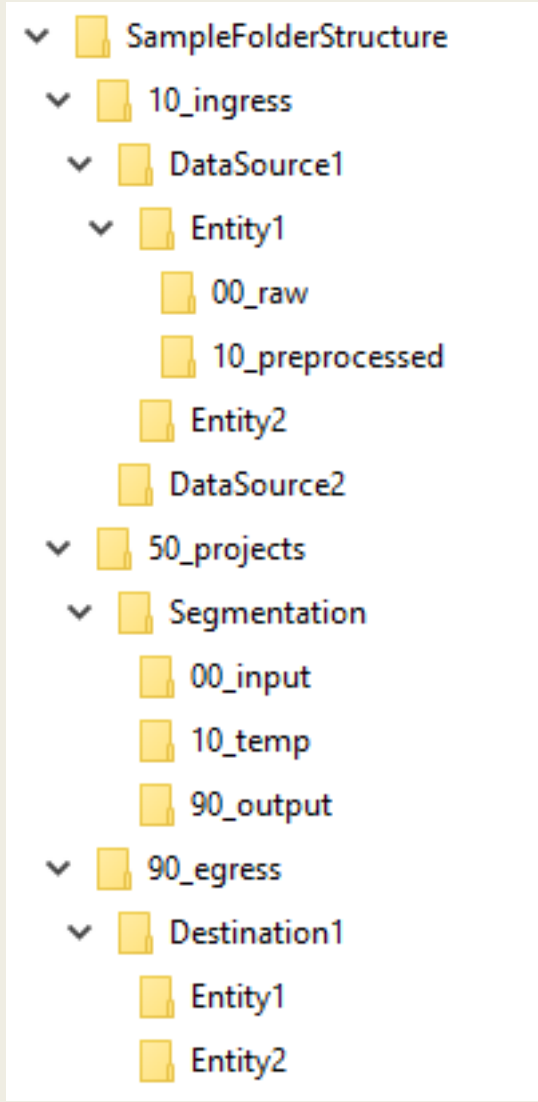
Also Consider:
- Multiple Tenants
- Versioned Sources



Getting Started - Storage

Security (ADLS only)!

- POSIX security – no inheritance
- Default permissions
- Use AAD Groups!
- Setup BEFORE loading data!

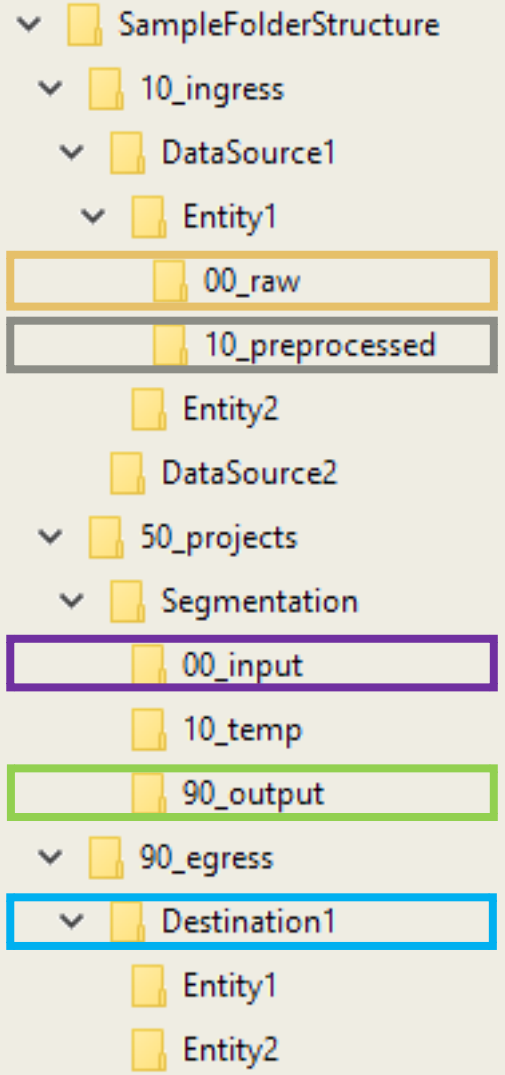


Getting Started - Storage

Clear definition of what is stored where!

Clear definition of transformation steps!

- 00_raw = 1:1 copy of source
- 10_preprocessed = 00_raw in common format
  no business logics here
- Projects input = filter/sub-select of 10_preprocessed
- Projects output = results after processing Projects input
- Egress = interface for consumers



Getting Started - Processing

Depends on the Use-Case!

All processing engines can be started on-demand!

Storage is separated and can be attached at run-time!

Apache Spark is defacto standard

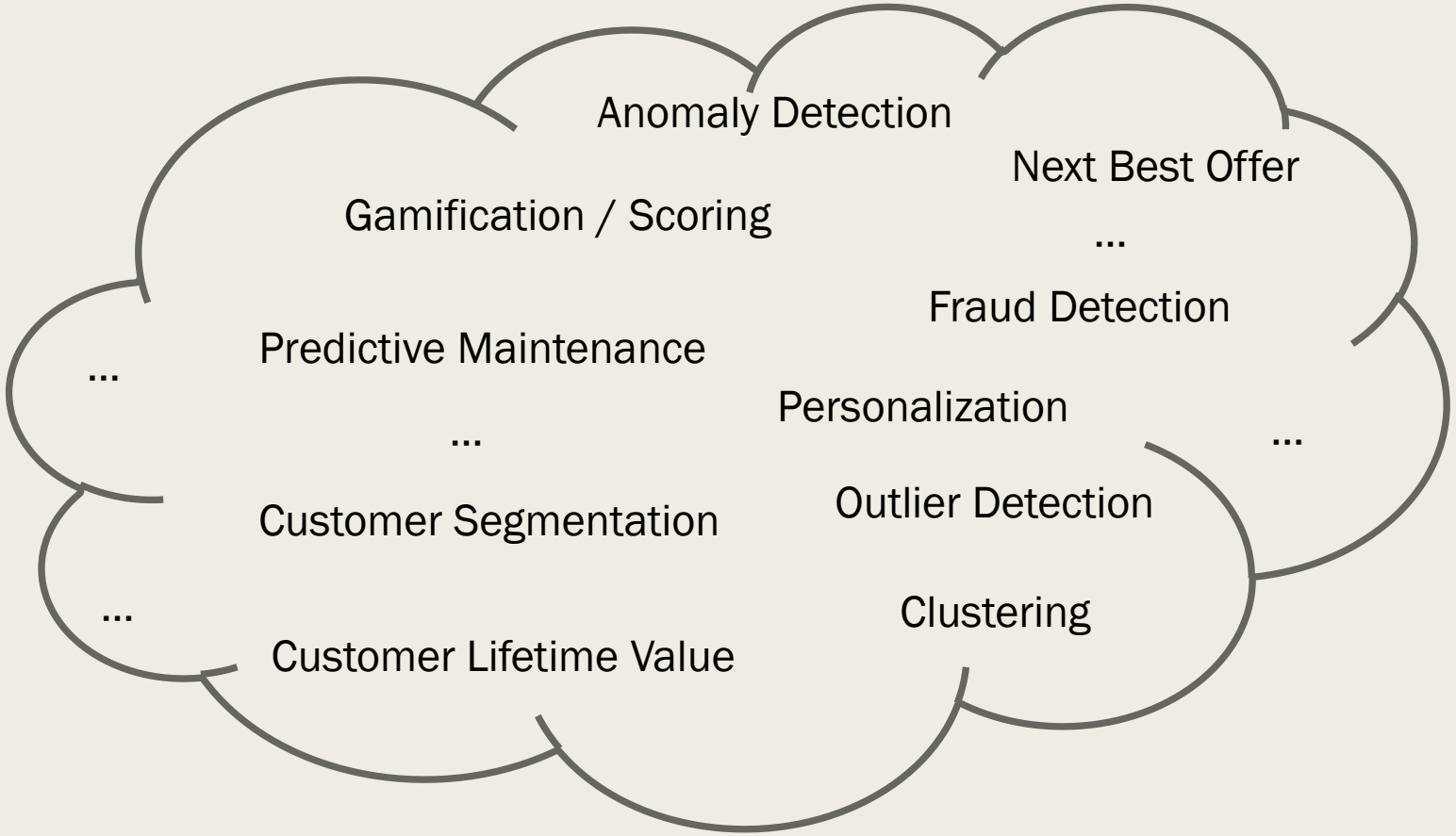
Getting Started - Serving

On-Demand Serving

- API / Microservice
- Streaming

Batch Processing

- scheduled run

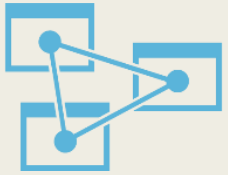


Getting Started - Serving

SQL

- Azure Synapse not designed for ad-hoc queries
- Use SQL DB datamarts!

Hive / SparkSQL



OLAP

- Power BI
- Azure Analysis Services

Getting Started – Data Management

Data Factory

- Manage your data movements
- Built-in Copy feature
- Orchestration of other services

Azure Purview

- Catalogize your data
- Make it searchable
- Apache Atlas API integration
- (Lineage analysis)

Scenarios

- Customer Analytics
  - *Customer Profiling / Customer 360°*
  - *Customer Online Journey*
- Advanced Analytics on sensor data from mobile app
  - *Telematics, GPS, …*
- Modern DWH for Analytics
- Data Screening

Customer Analytics

Scenario

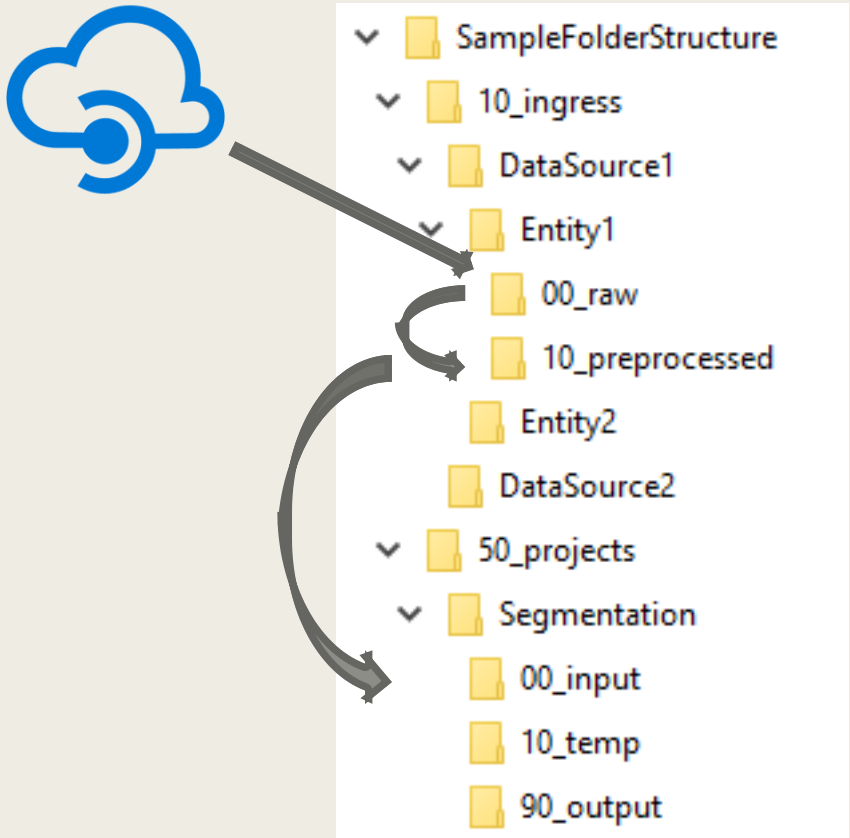
- Customer Online Journey, Customer 360°
- SEO/SEA, Newsletter, Banner, …
- Web logs (1000+ columns)
- Various APIs (Google, Exactag, …)
- Hourly loading

Use AI and ML to better understand and target the customer

Customer Analytics

- Data ingestion via
  ADF pipelines + Databricks Notebooks
  - *REST APIs*
  - *Web Logs*
  - *CRM / SQL*
  - *Web Shop*

- Various use-cases
  - *Customer Lifetime Value calculation (batch)*
  - *Segmentation (batch)*
  - *Next Best Offer (on-demand, web shop)*



Advanced Analytics on sensor data from mobile app

Scenario

- Score driving behavior of users
- Mobile App tracks movements
- Azure IoT infrastructure
- 100k+ devices
- HERE Maps enrichment
- White-labeled solution

Use ML to analyze driving behavior, detect maneuvers, concentration level, …

Advanced Analytics on sensor data from mobile app

- Interface to IoT platform is EventHub Capture
  - *AVRO files with compressed, nested JSON (ZLib + MsgPack)*
- Versioned files/messages (schema changes)
- Hourly extraction
- Daily processing
  - *HERE Enrichment*

Extensive use of Databricks

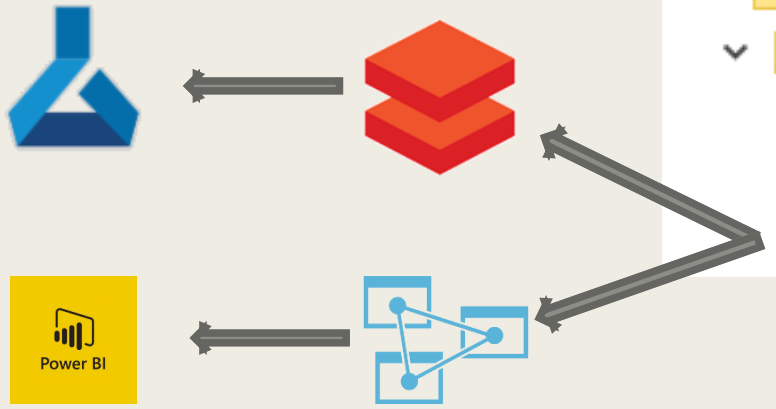
- Very flexible
- Very scalable

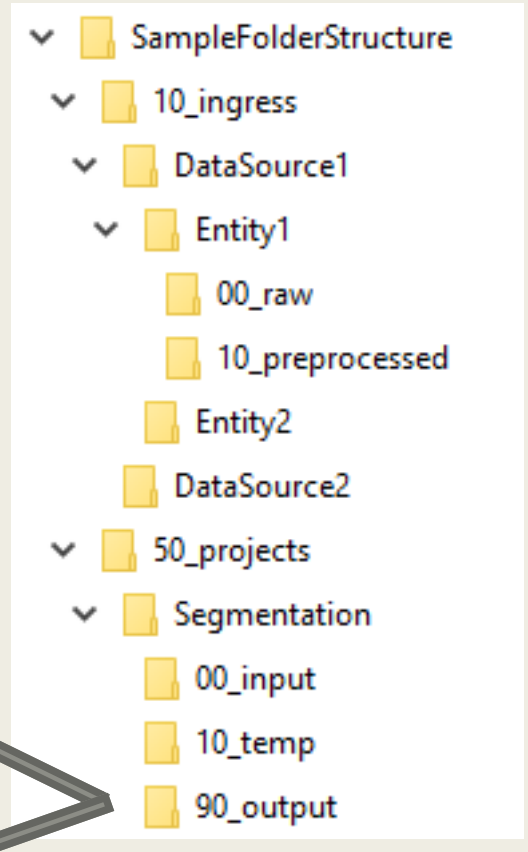
Advanced Analytics on sensor data from mobile app

Separate "Projects" for different Use-Cases

- *Analytical SQL model for Data Scientists*
- *Reporting*

- ML model development/training on historic data
- High-Level reporting





Modern DWH

Scenario

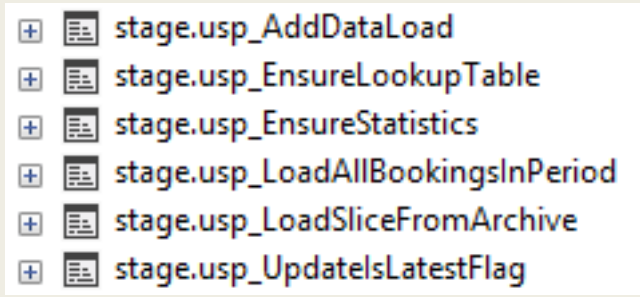
- Online Booking System (24/7)
- 500 GB database
- Multiple Target Systems to load (SQL, API, JSON,
- Load should have little impact on live system
- Only last 18 months in live system
- Rebuild full history any time (>18 months)
- Reporting in Cloud with Azure Analysis Services + PowerBI

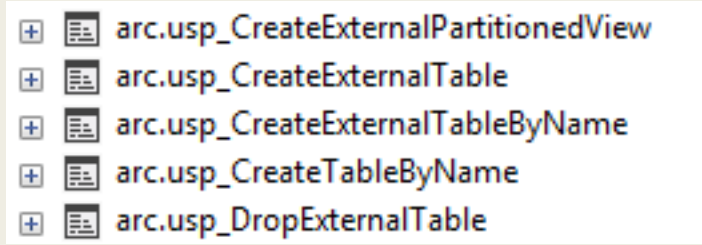
Data Lake as DWH Staging Area/Archive

Solution

- Azure Data Factory pipelines
  - *Nightly export of data (createdOn/modifiedOn)*
  - *Copy SQL to ORC files*
  - *Create SQL Synapse external table for last day dynamically*
  - *Populate temporary staging tables*

- Trigger regular DWH ETL using SSIS in VM





Data Screening

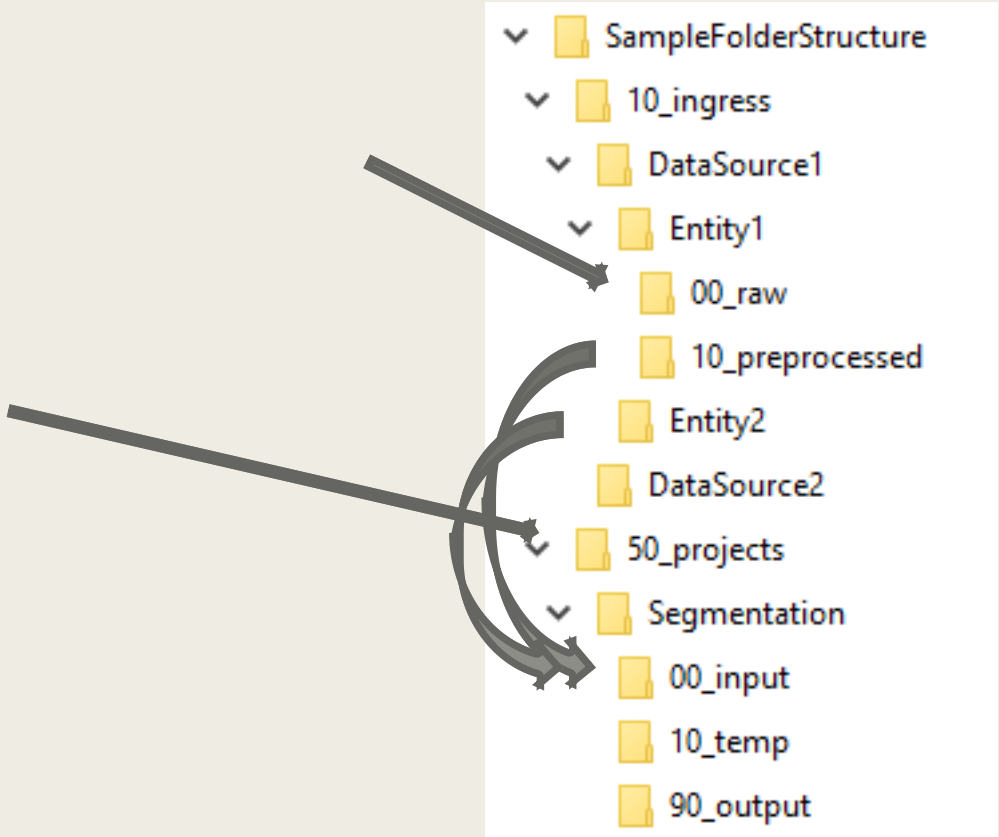
Scenario

- New data source is discovered
- Need to analyze for valuable insights
- Check possibilities to combine it with existing data

Data Screening

Solution

- Load data into ADLS (one-time load to /ingress)
- Create project
  - *Also add data from other sources*
- Process data to get a structured format
  - *Using existing tools (DataBricks, Synapse, …)*
- (try to) combine with reference data
  - *From any other source/project*
- Analyze for insights



Take-Aways

- Storage is the key – plan it well!
- Can be beneficial for many use-cases (DWH, AA, …)
  - *Favors Advanced Analytics and Machine Learning*
- Stick to your design/architecture
  - *Go with PaaS Services if possible*
  - *Keep used technologies at a minimum*

File Layout & Formats

- Query oriented vs. ETL oriented
  - *Partitioning*
  - *Compression*
  - *Number of files*

- File Format
  - *Delta*
  - *Parquet*
  - *JSON*

Thanks!

@gbrueckl

blog.gbrueckl.at

gerhard@gbrueckl.at

https://github.com/gbrueckl

DatabricksPS

Databricks VSCode

paiqo

www.paiqo.com

SQL saturday VIENNA

PASS AUSTRIA