



Herzlich Willkommen bei ORAYLIS

Komplettanbieter für individuelle und nutzerorientierte
Business Intelligence-, Data Analytics- und Artificial Intelligence- Lösungen

**Vom
Modern Data Warehouse (MDWH)
Zu
Synapse**

Jens Kröhnert – Principal Consultant
Daniel Esser – Senior Solution Architect

Wir sind ORAYLIS



115 MITARBEITER

davon 90 Experten, die für BI,
Data Analytics & AI brennen



21 JAHRE

prall gefüllter
Erfahrungsschatz



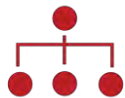
15 JAHRE

Microsoft Partner mit
4 Gold-Kompetenzen



800+ PROJEKTE

in den Bereichen BI, Data Analytics
& AI erfolgreich umgesetzt



100 PROZENT

inhabergeführt und durch
flache Hierarchien geprägt



20 PROZENT

Wachstum pro Jahr



100 Prozent

Weiterempfehlung
bei kununu



3 AUSZEICHNUNGEN

Wir sind ein
Great Place to Work

Das machen WIR

DATEN. INFORMATIONEN. MEHRWERT.

Wir realisieren unter Anwendung agiler Projektmethoden exklusive Business Intelligence-, Data Analytics- und Artificial Intelligence- Lösungen – nicht von der Stange, sondern maßgeschneidert für Ihre Bedürfnisse.

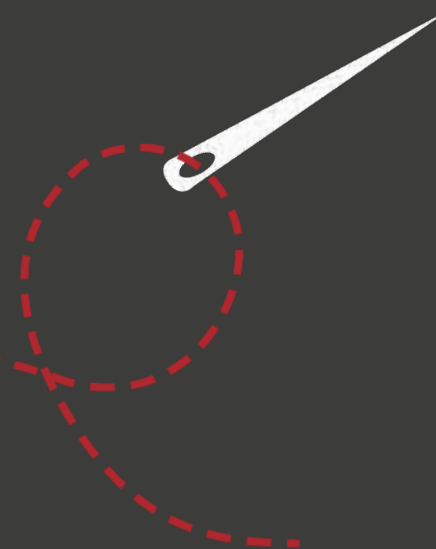
DATENSTRATEGIE
& KONZEPT



IMPLEMENTIERUNG
& AUFBAU



BETRIEB
& OPTIMIERUNG



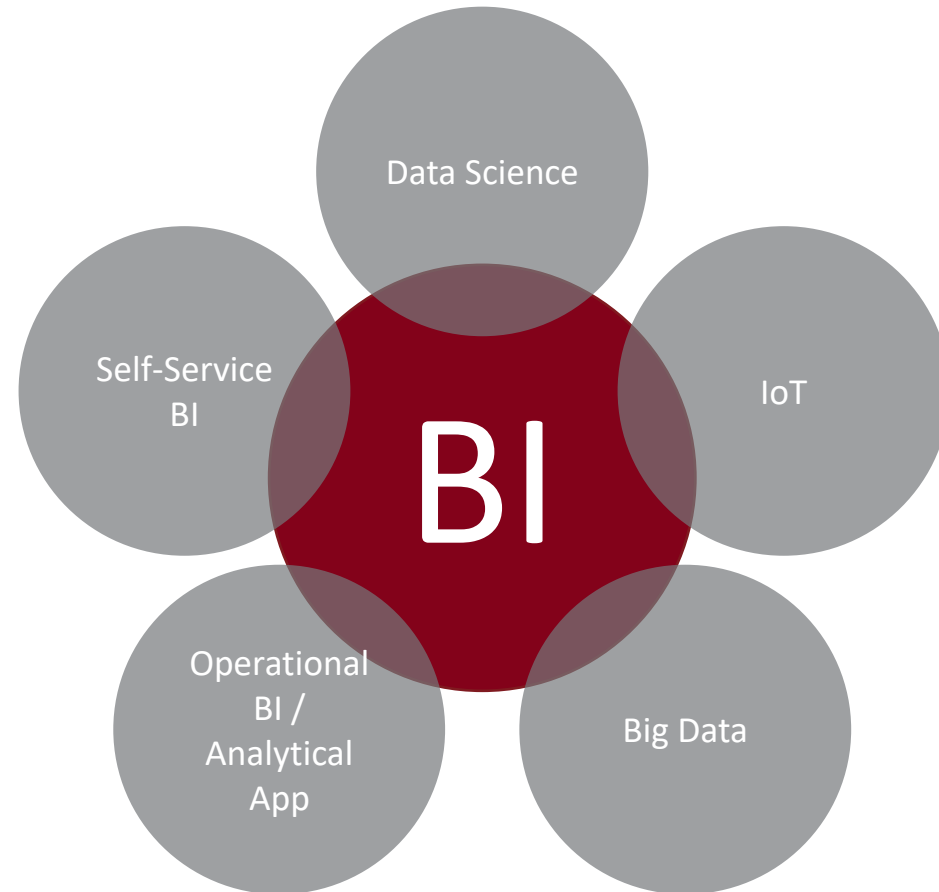
We turn
DATA
INTO
VALUE

Das können WIR



BI++ auf Azure = MDWH

BI – Basis für Digitalisierung

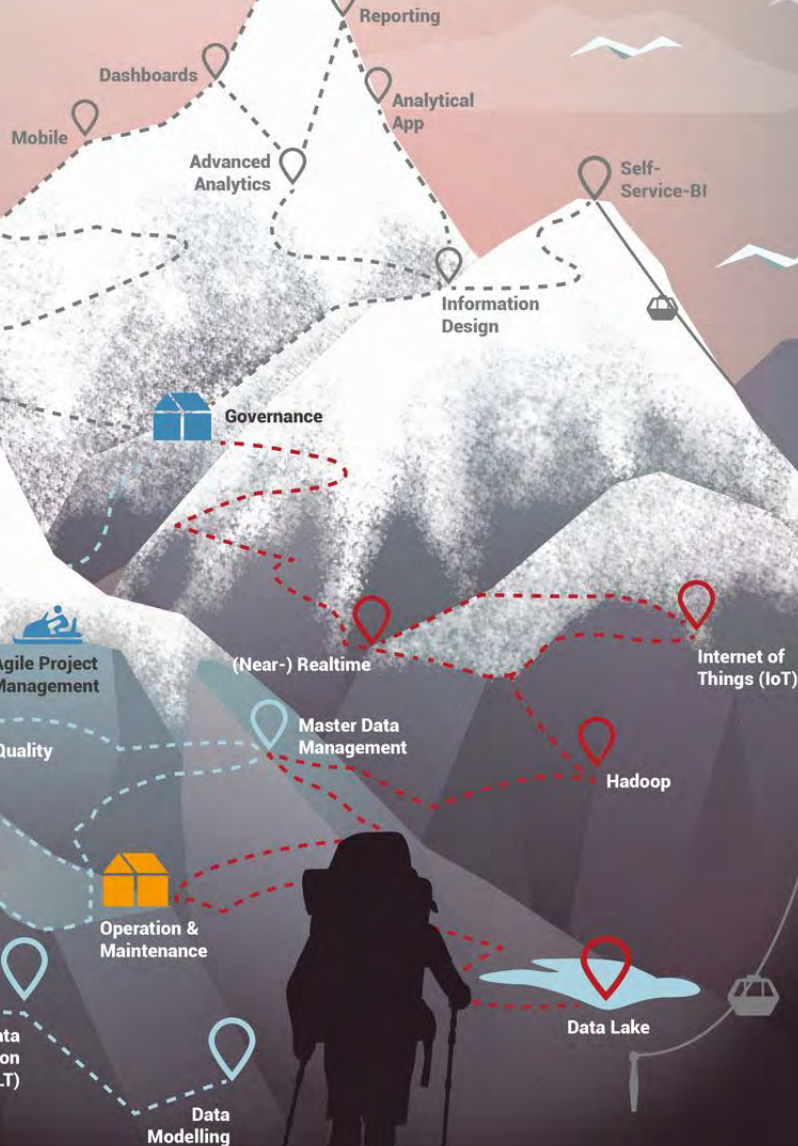


Data Analytics Platform

- TRAIL -

360° View

Data Driven Innovation



Unsere Leitsätze

Simplicity - Die Plattform muss einfach sein. Das fördert die aktive Nutzung und Akzeptanz im Unternehmen. Gleichzeitig reduziert sich der Schulungsaufwand.

Speed - Die Plattform muss schnell sein – von der Umsetzung über die Datenbereitstellung bis zur Analyse. Je früher die gewünschten Informationen bereitstehen, desto höher ist der Nutzen.

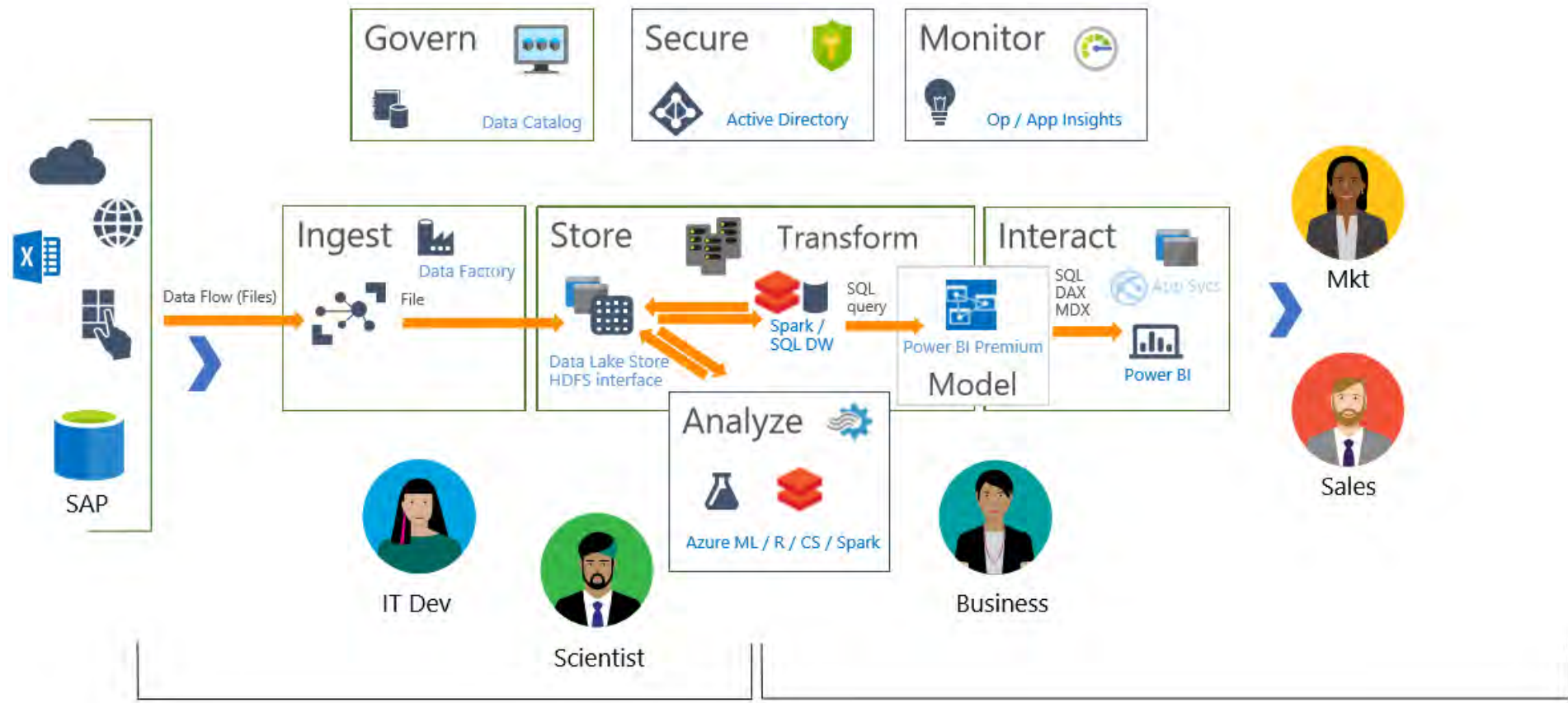
Data Quality - Die Daten müssen konsistent, vollständig, glaubwürdig und aktuell sein. Werden diese Erwartungen nicht erfüllt, stößt die Lösung auf Ablehnung.

Ability to Change - Die Plattform muss verändert werden können – schnell, stabil, bezahlbar. Ohne diese Fähigkeit steht die Lösung nach kurzer Zeit auf dem Abstellgleis.

Agility - Die Plattform muss flexibel und dynamisch entwickelt werden. Die Komplexität von Projekten lässt sich niemals vollständig vorhersehen. Agilität steigert die Qualität und verringert die Kosten.

Speichern und Verarbeiten jeglicher Daten und Formate

MODERN DATA WAREHOUSE



Kernkonzept MDWH on Azure

Trennung von Storage und Compute

Big Data Use Cases

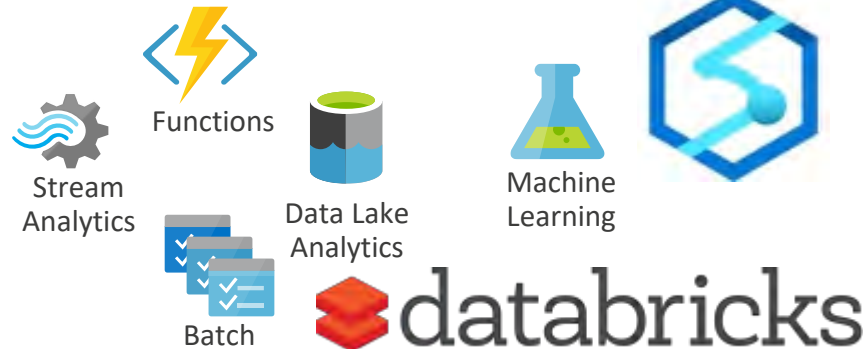
Ingest & ETL



Streaming



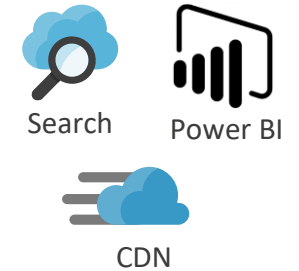
Analytics & Machine Learning



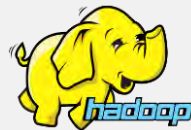
Data Aggregation



Presentation



Azure HDInsight



Blob Storage



Data Lake Store

Vergleich

DATA LAKE & DATA WAREHOUSE

Data Lake

Beliebige Datenstrukturen und Formate

Streaming, micro-batch oder Batch-Prozesse

ELT
(Extract, Load und Transform zum Zeitpunkt
der Datenverarbeitung)

Daten beschaffen, analysieren und durch ein
iteratives Vorgehen die beste Struktur ableiten

Zum Zeitpunkt der Abfrage: "Schema-on-read"
oder wenn "curated" Daten erzeugt werden

Datenformat

Typischer Datenladeprozess

**Typisches
Verarbeitungsschema**

Vorgehen zur Analyse

**Wann erhalten Daten
eine Struktur?**

Data Warehouse

strukturierte Daten in Spalten und Zeilen einer
relationalen Datenbank

Batch-Prozesse

ETL
(Extract, Transform, then Load)

Datenstruktur definieren, Daten beschaffen, dann
analysieren. Iterativ die Datenstruktur an den
Bedarf anpassen.

Zum Zeitpunkt der Datenverarbeitung: "Schema-
on-write"

Vergleich

DATA LAKE ONPREM VS CLOUD

Data Lake OnPrem

Limitierte Ressource (Teuer)

Kostet nur Strom (Billig)

Rohdaten (RL) wird persistiert

Storage

Compute

Typisches
Verarbeitungsschema

Data Lake Azure

Unlimitierte Ressource (Günstig)

Kosten nach Nutzung (Teuer)

Rohdaten (RL) UND weitere Layer (TL, BL) werden persistiert

Varianten Speicher auf Azure für Data Lake Storage?

Blob Store (v2)

- General purpose object store
- Containers -> Blobs
- Non Hierarchical Store
- HDFS based

Azure Data Lake Store Gen1

- Hierarchical file system
- Folder -> Files
- Non Hierarchical Store
- HDFS equivalent

Azure Data Lake Store Gen2

- Best of both worlds
- Hierarchical file system API
- Blob Store API
- Blobstore v2 + Hierarchical Namespace

Azure Blob Storage : Object Store

 /TaxiData/RawData/Yellow/Yellow_Taxi_Trip_Data_2018.csv
 /TaxiData/RawData/Yellow/Yellow_Taxi_Trip_Data_2019.csv

Azure Data Lake Gen 1: Hierarchical Storage

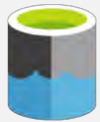
 /TaxiData
  /RawData
  /Yellow
  /Yellow_Taxi_Trip_Data_2019.csv

Which main services do WE USE AND WHY

AZURE

DATA LAKE STORAGE

- ✓ Highly scalable and cost-effective data lake service for big data analytics.
- ✓ It combines the power of a high-performance file system with massive scale and economy.



AZURE

DATA FACTORY

- ✓ Delivers hybrid data integration service that simplifies Extract-and-Load at scale.
- ✓ No code or maintenance required.



AZURE

DATABRICKS

- ✓ Unlocks insights from all your data with artificial intelligence (AI) and data processing power at scale.
- ✓ Based on Spark which is an open-source general-purpose cluster-computing framework, no vendor lock-in, supported by a big community.



MICROSOFT

POWER BI

- ✓ Business analytics service which delivers insights to enable fast, informed decisions.
- ✓ It seamlessly integrates into custom B2B solutions and has one of the best in class analytics engines.



Data Warehouse Layer Architecture

Einblicke in die Datenarchitektur

SL – Serving Layer

BL – Business Layer

TL – Technology Layer

RL – Replication Layer



Ob Namen mit Zonen oder Layern – es braucht eine gute... Schichtenarchitektur



Praxis-Projekt Daten

Adventure Works Cycles

fiktiver Fahrradhersteller und -verkäufer

Verkauf über Internet-Shop und Reseller

Produkte

Bikes
Accessories
Clothing
Components

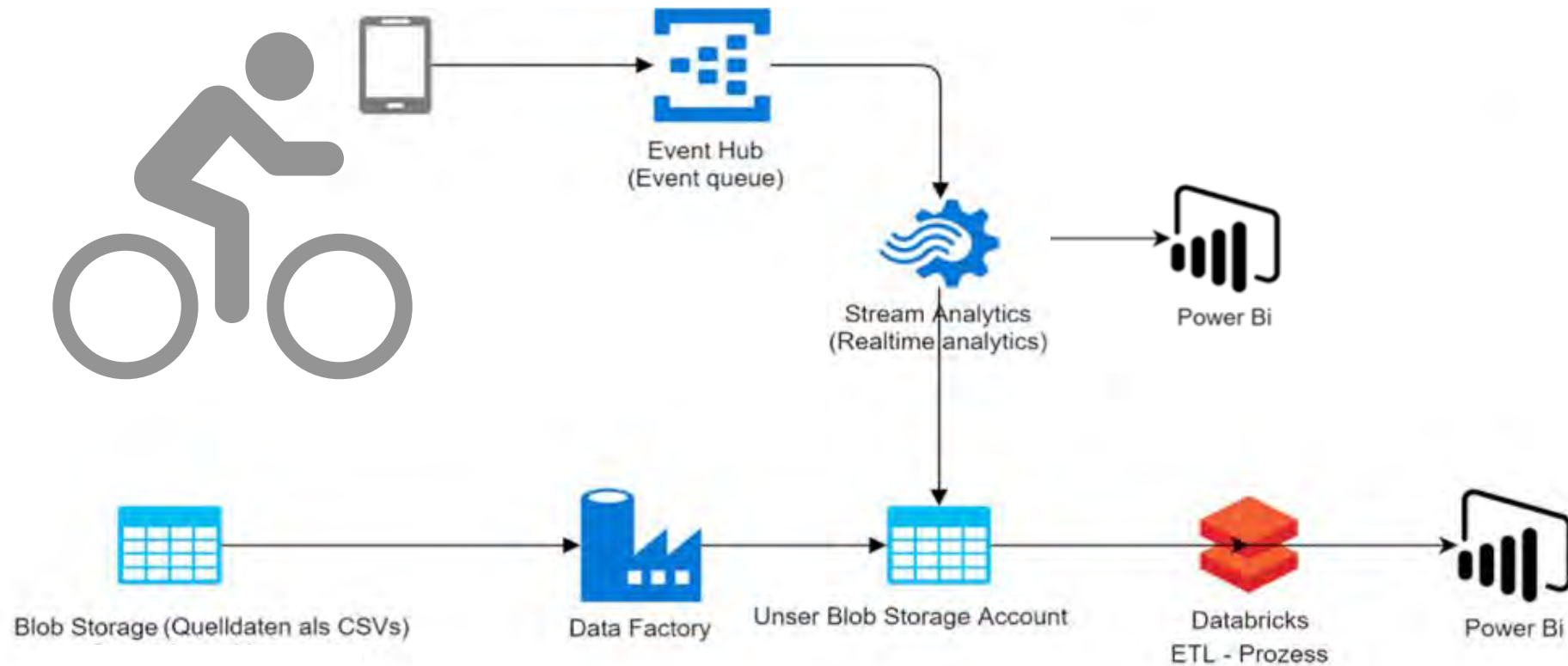
Multinational



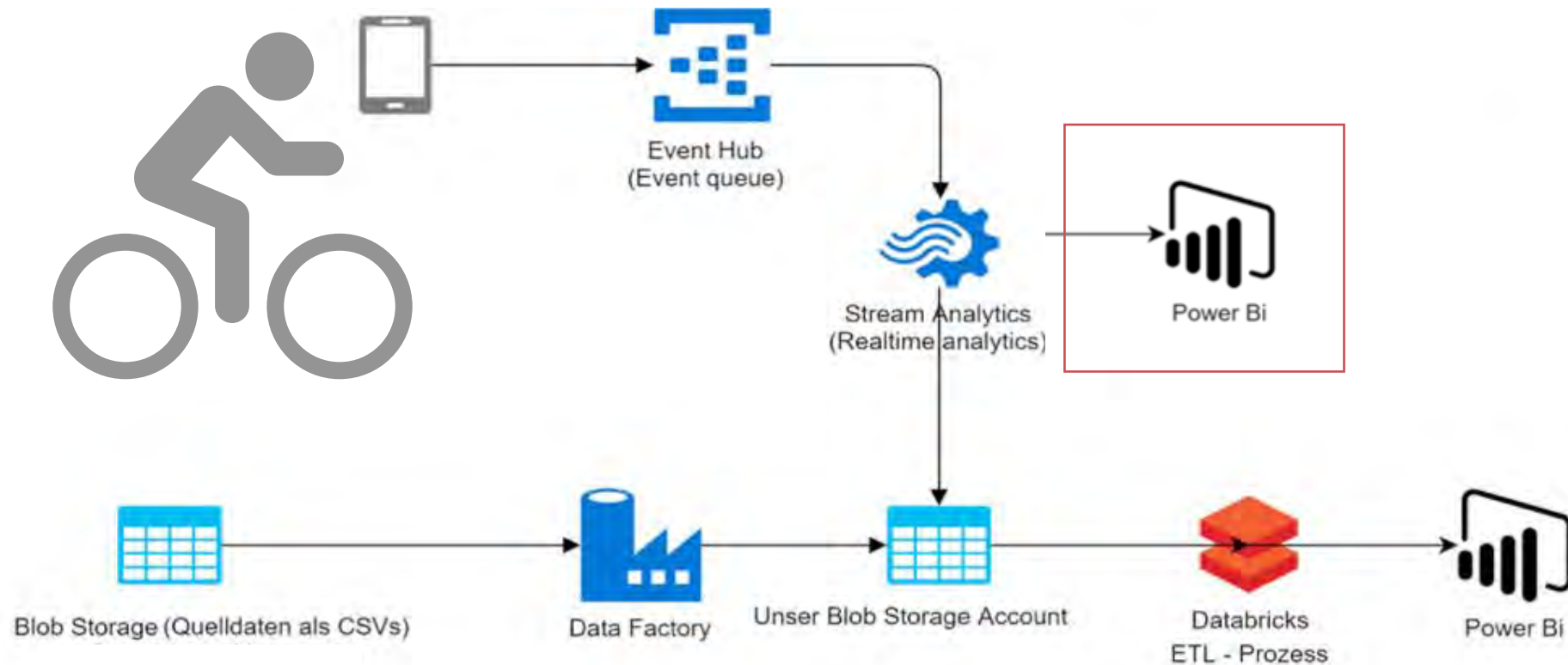
Name
[..]
DimCustomer.csv
DimDate.csv
DimProduct.csv
DimPromotion.csv
DimReseller.csv
DimSalesTerritory.csv
FactInternetSales.csv
FactResellerSales.csv

		Dimensions											
			Order Date	Ship Date	Delivery Date	Customer	Reseller	Promotion	Product	Sales Territory	Posting Date		
Facts	Priorität	Σ	2	2	2	2	2	3	3	3	1	0	
Internet Sales	1	7	1	1	1	1		1	1	1			
Reseller Sales	1	7	1	1	1		1	1	1	1			

Praxis-Projekt MDWH Architektur



Praxis-Projekt MDWH Architektur

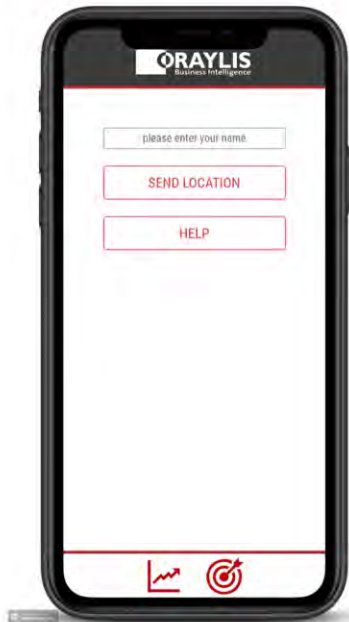


Let's go for a virtual bicycle ride!

Shake Challenge!

IoT-Demo with your smartphone

Evaluate your sensor data in realtime with Power BI



1. Open [https://](https://azurewebsites.net/) azurewebsites.net/
2. Press "HELP" and check for further requirements
3. Enter your name and press "SEND LOCATION"
4. SHAKE!
5. Press "STOP SENDING" and view your results in Power BI under:  and 

Bitte als
Benutzernamen
eine Zahl zwischen
0 und 12345
wählen

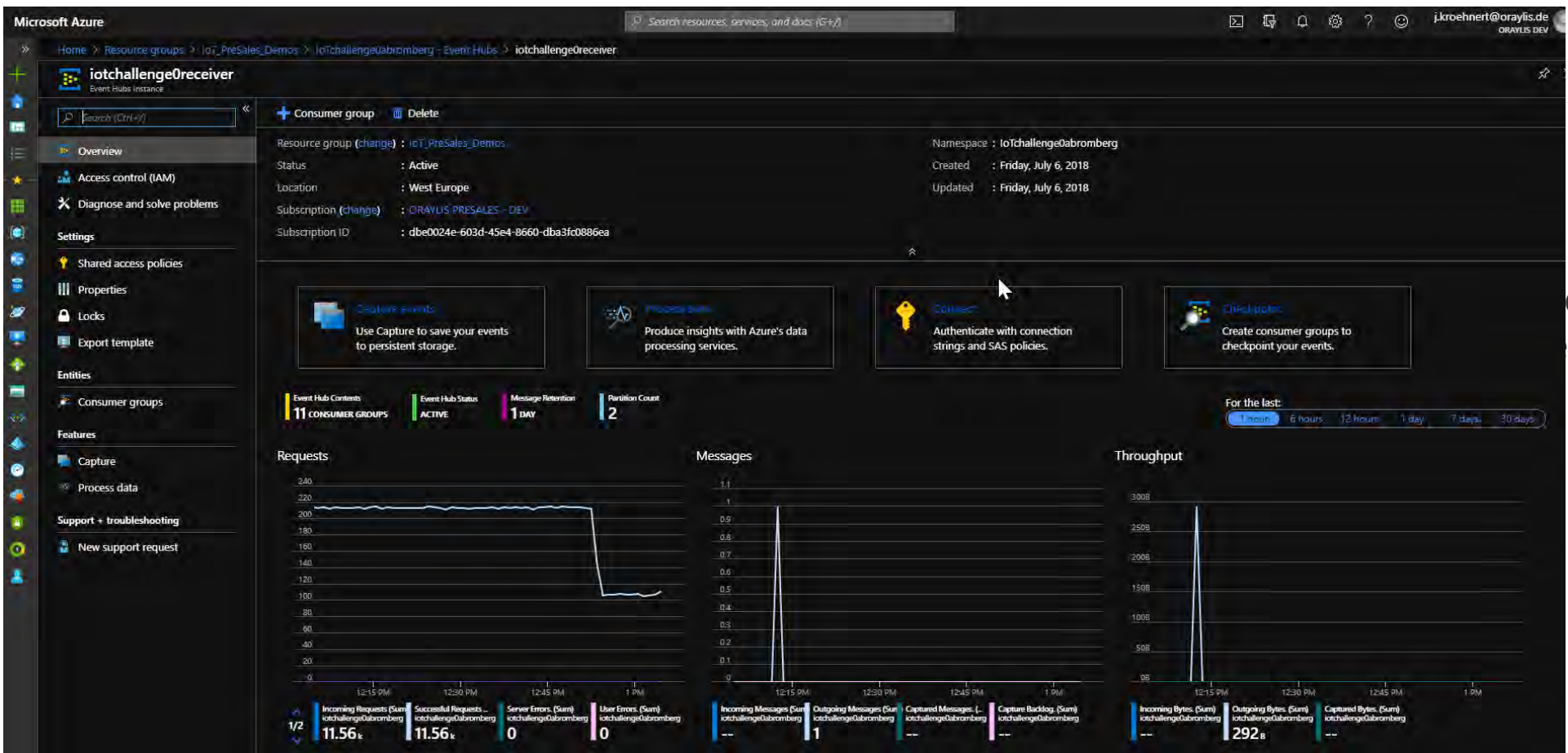
Gehen Sie auf <http://.azurewebsites.net/> und schütteln Sie ! 😊

Azure Stream Analytics Input Event Hub



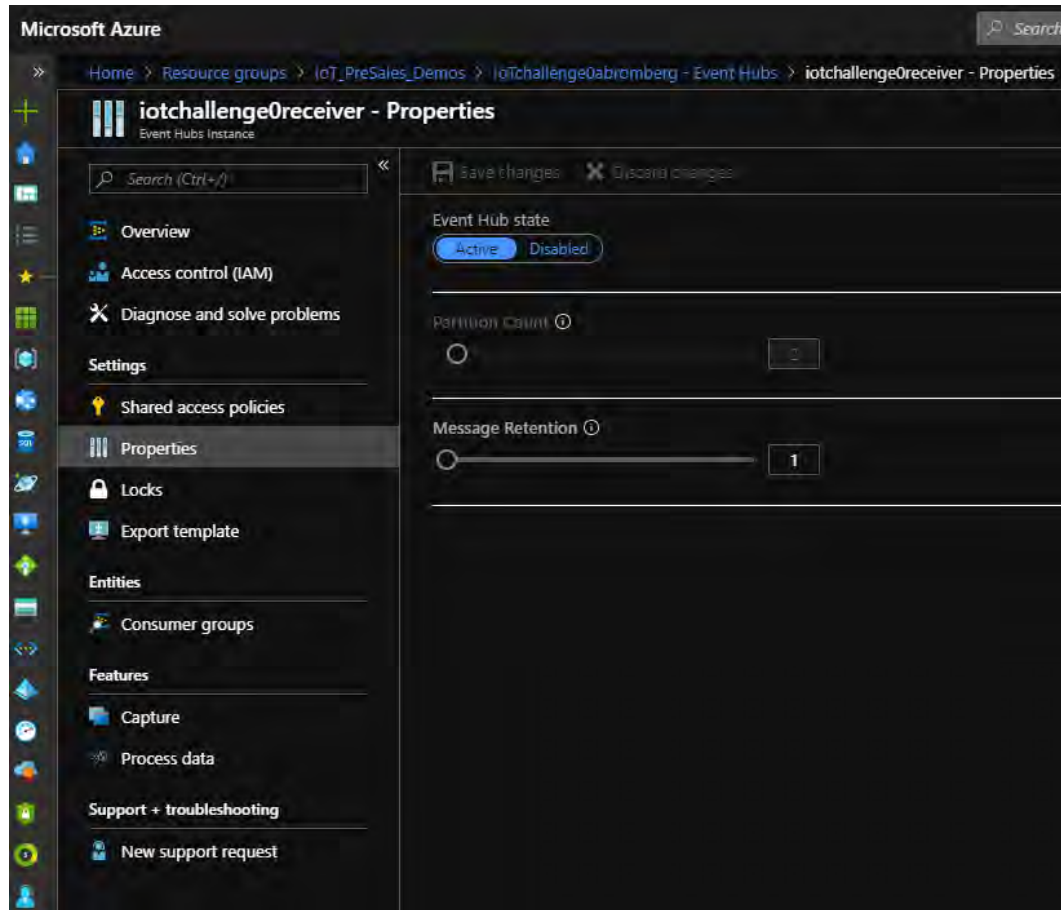
Azure Event Hub, der kleine Bruder der Azure IoT Hubs kann nur Daten unidirektional entgegennehmen über verschiedene Protokolle (SSL, AMQP, MQTT). Es ist sozusagen ein Endpunkt in der Cloud mit angehangener Queue als Speicher für Abnehmer (Consumer-Groups)

Azure Event Hub

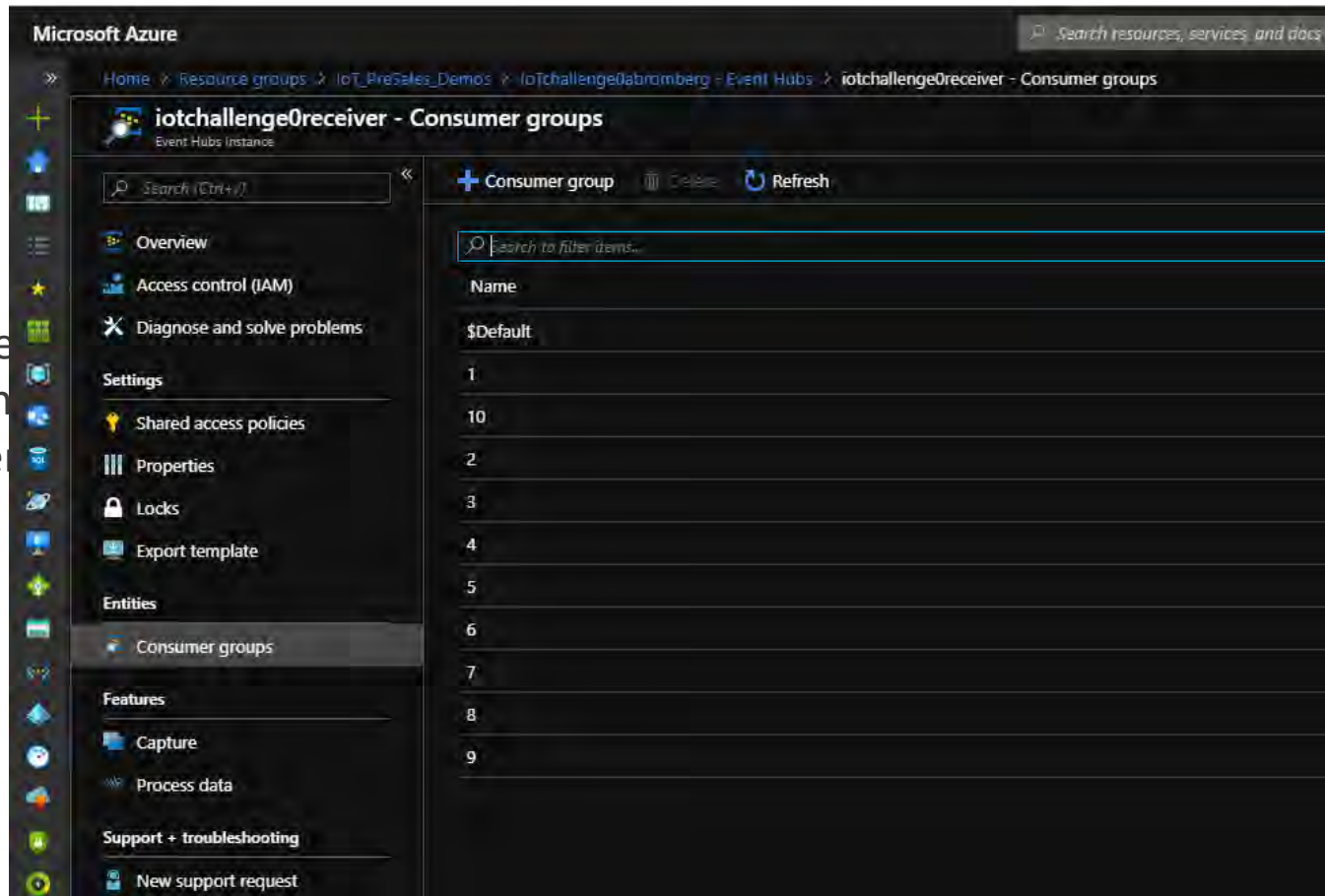


en
rden

Azure Event Hub



Intel
In
ver



Azure Stream Analytics

Start Stop Delete

Resource group (change) : BI-BasisV2-001 Send feedback : UserVoice

Status : Starting Created : Wednesday, August 28, 2019, 9:21:55 AM

Location : West Europe Started : Wednesday, August 28, 2019, 10:05:02 AM

Subscription (change) : ORAYLIS Intern - Schulungen Output watermark : -

Subscription ID : 04154975-918c-4f54-b216-f8ca3cdee657 Hosting environment : Cloud

Inputs

1 eventHub Event Hub

Outputs

2 BlobStorage Blob storage

PowerBI Power BI

Query

```
1 WITH
2 FULLQUERY AS
3 (
4 SELECT
5 EVENTENQUEUEDUTCTIME,
6 DATEADD (HH, 2, EVENTPROCESSEDUTCTIME) AS TIMESTAMP,
7 EVENTPROCESSEDUTCTIME,
8 LONGITUDE,
9 LATITUDE,
10 USERNAME,
11 IP,
12 OS,
13 VOTE,
14 SHAKE,
```

basisv200010storage

Storage account key

Subscription

Subscription information not needed

* Container

Create new Use existing

datalake

* Service Bus namespace

ioTchallenge0abromberg

* Event Hub name

Path pattern

logs/(date)/(time)

* Output alias

PowerBI

Date format

YYYY/MM/DD

Group workspace

My workspace

Time format

HH

* Dataset name

BasisV2001

* Event serialization format

CSV

* Table name

BasisVS2001

Currently authorized as [Alexander Bromberg](#)
(a.bromberg@oraylis.de)

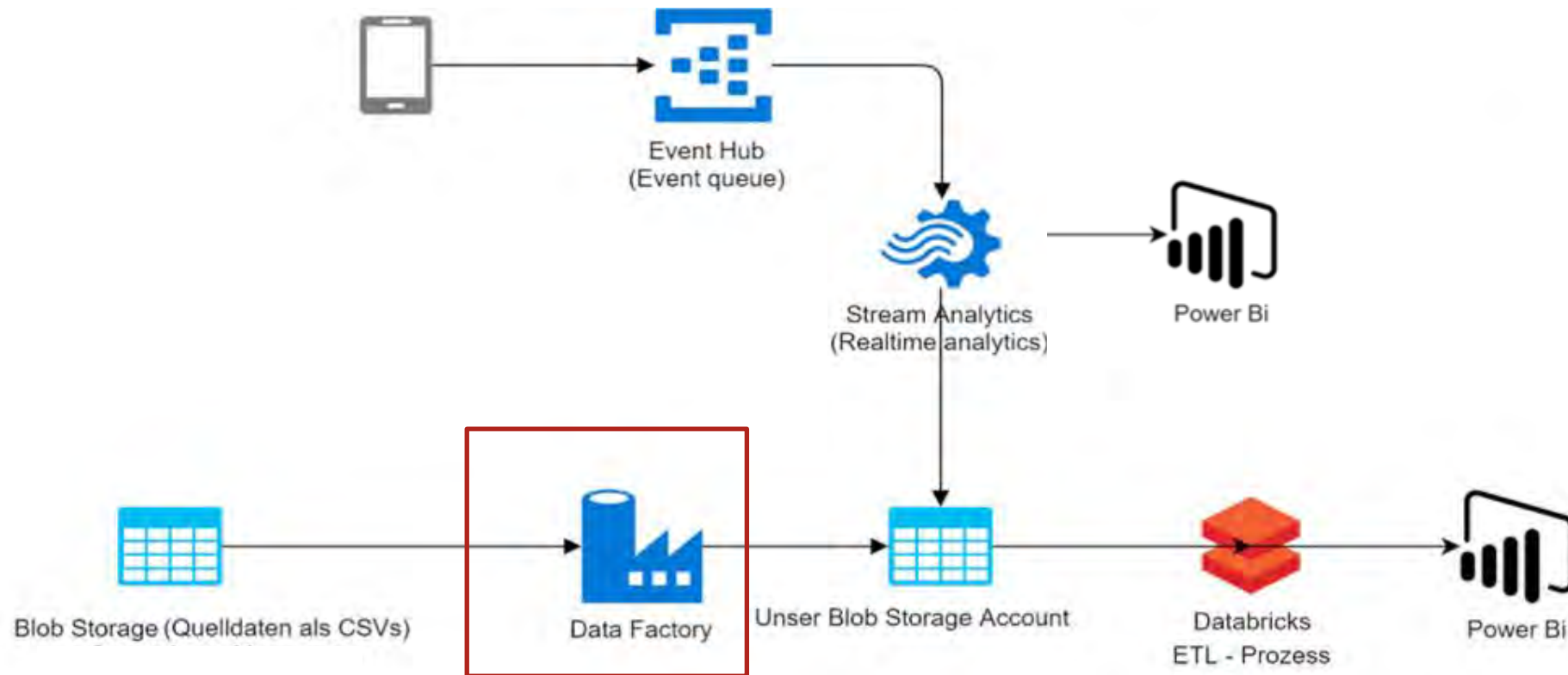
Authorize connection

You'll need to authorize with Power BI to configure your output settings.

Authorize

Erstellung von Stream Analytics sowie die dazugehörigen Inputs, Outputs und die SQL Query

Praxis-Projekt MDWH Architektur



Azure Data Factory

The screenshot displays the 'Edit Linked Service' dialog for 'BI-BasisV2001Storage' in the Azure Data Factory portal. The dialog is divided into two main sections: 'General' and 'Connection'. The 'General' section includes fields for 'Name', 'Description', 'Connect via integration runtime', 'Authentication method', 'Account selection method', 'Account selection method', 'URL', and 'Storage account key'. The 'Connection' section includes fields for 'Linked service', 'File path', 'Compression type', 'Column delimiter', and 'Row delimiter'. The 'File path' field is highlighted with a blue box, showing a dynamic path expression: `@CONCAT('RL/DimProduct/snapshot_date',SUBSTRING(formatDateTime(utcnow(),'u'),0,19))`. The 'Compression type' is set to 'none', 'Column delimiter' is 'Comma (,)', and 'Row delimiter' is 'Auto detect (\r,\n, or \r\n)'. The 'Test connection' button is visible, and a green checkmark indicates 'Connection successful'.

Edit Linked Service (Azure Data Lake Storage Gen2)

Name *

Description

Connect via integration runtime *

Authentication method

Account selection method

URL *

Storage account key *

Annotations

File path *

Compression type

Column delimiter

Row delimiter

Test connection

Cancel

Finish

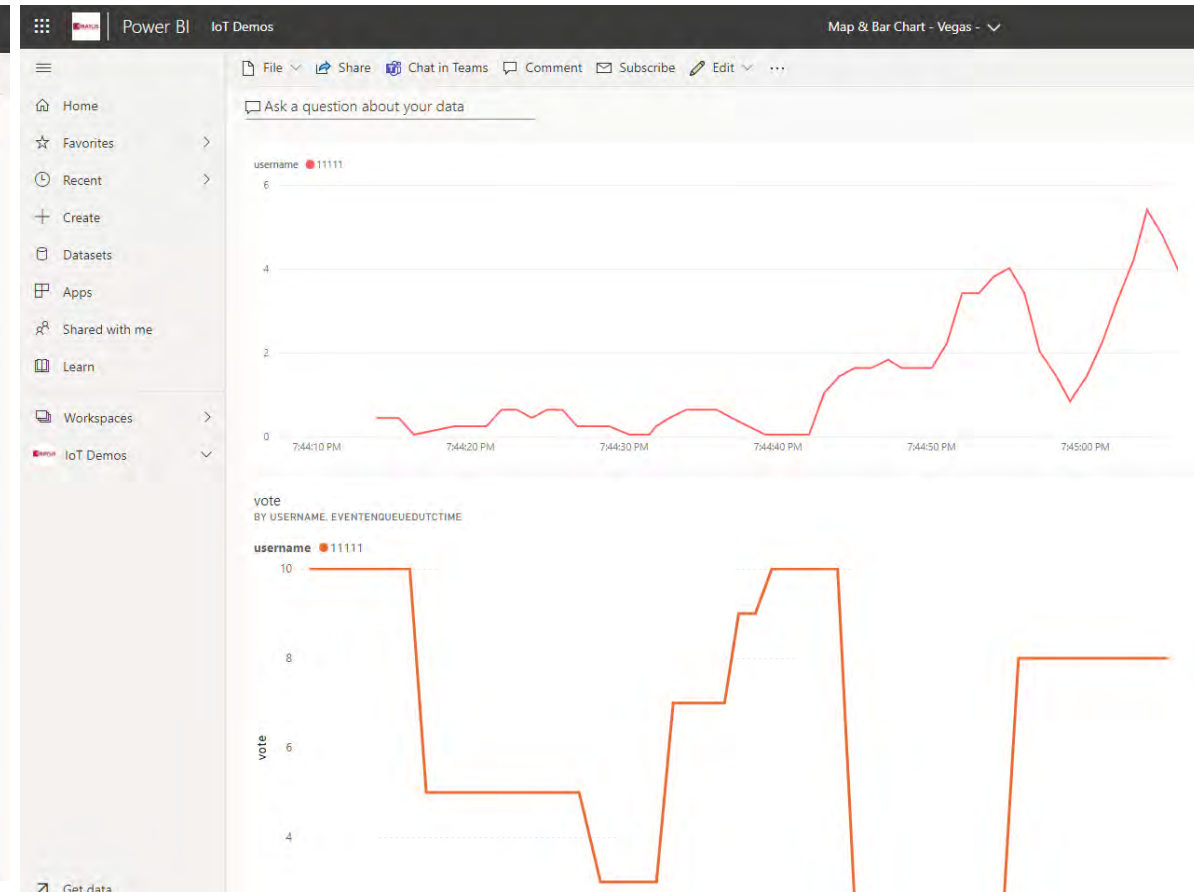
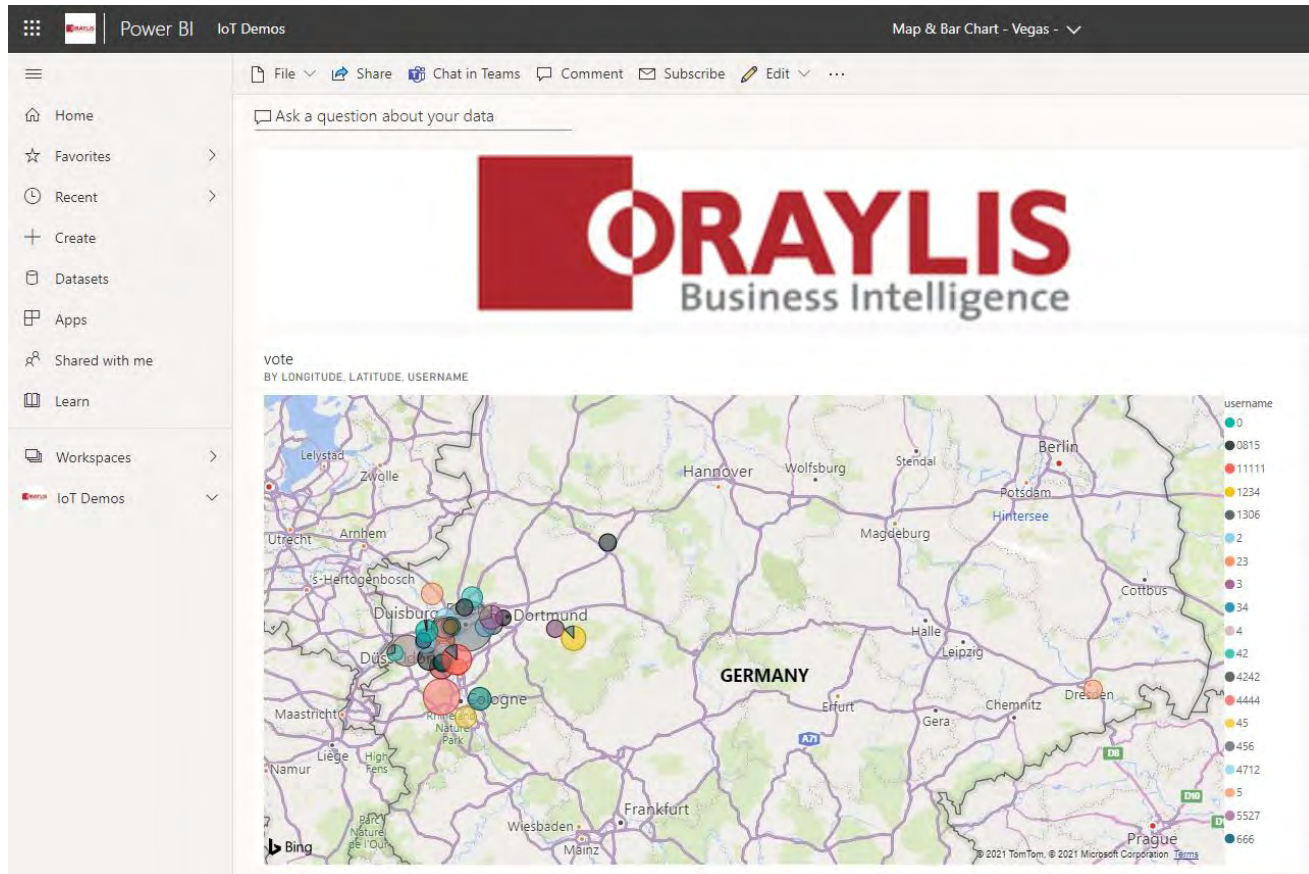
Dynamic Path beim Ziel-Dataset für den Copy Prozess

Azure Storage Explorer

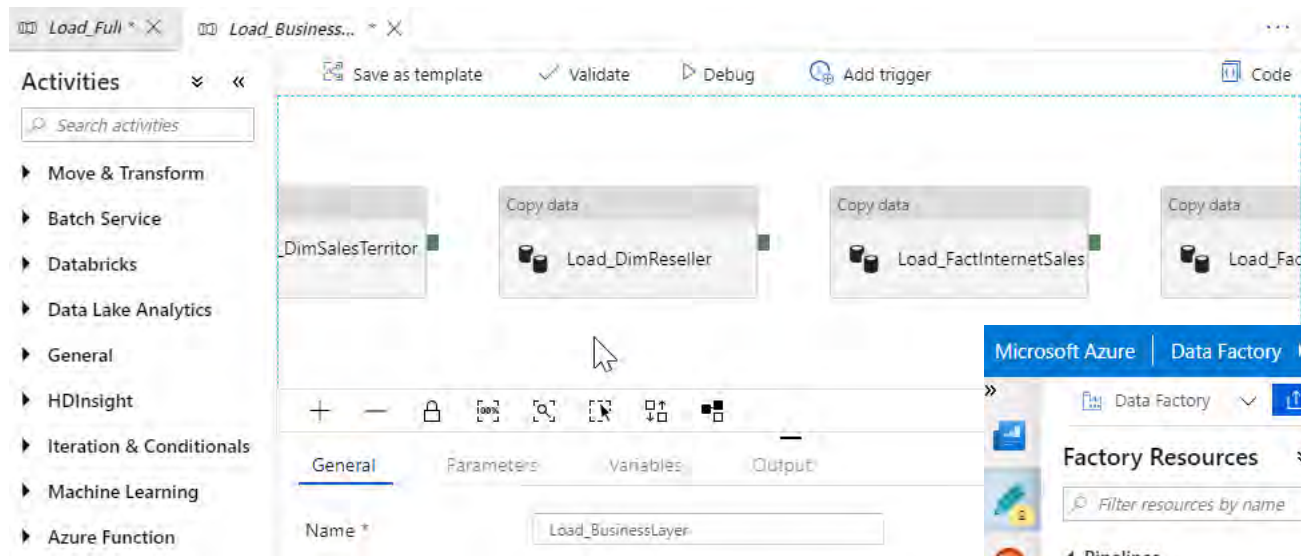
The screenshot displays the Microsoft Azure Storage Explorer application. The top navigation bar shows the path: Home > BI-BasisV2-Master > bibasisv20master. The main title is "bibasisv20master | Storage Explorer (preview)". The left sidebar contains a search bar and a list of navigation options: Overview, Activity log, Tags, Diagnose and solve problems, Access Control (IAM), Data transfer, Events, and Storage Explorer (preview) (which is selected). Below these is a "Settings" section. The central pane shows a tree view of "BLOB CONTAINERS" with folders: aas, assets, bootdiagnostics-masterdsv-d760d7, businesslayer, datalake (selected), and excelsource. The right pane shows a table of "Active blobs (default)" for the "datalake" container. The table has columns: NAME, ACCESS TIER, ACCESS TIER LAST MODIFIED, LAST MODIFIED, BLOB TYPE, CONTENT TYPE, SIZE, and STATUS. The table lists 15 folders, all of type "Folder", with names starting with "snapshot_date=" followed by a timestamp. A mouse cursor is hovering over the second row.

NAME	ACCESS TIER	ACCESS TIER LAST MODIFIED	LAST MODIFIED	BLOB TYPE	CONTENT TYPE	SIZE	STATUS
snapshot_date=2019-09-20 12:23:57				Folder			
snapshot_date=2019-09-26 10:02:58				Folder			
snapshot_date=2019-10-15 12:03:44				Folder			
snapshot_date=2019-10-16 21:54:26				Folder			
snapshot_date=2019-10-17 12:21:50				Folder			
snapshot_date=2019-12-04 13:00:33				Folder			
snapshot_date=2020-01-24 13:48:10				Folder			
snapshot_date=2020-02-05 08:27:16				Folder			
snapshot_date=2020-03-02 13:28:42				Folder			
snapshot_date=2020-03-04 11:01:13				Folder			
snapshot_date=2020-05-08 11:59:19				Folder			
snapshot_date=2020-05-08 12:17:49				Folder			
snapshot_date=2020-05-08 12:31:45				Folder			
snapshot_date=2020-06-30 11:18:00				Folder			

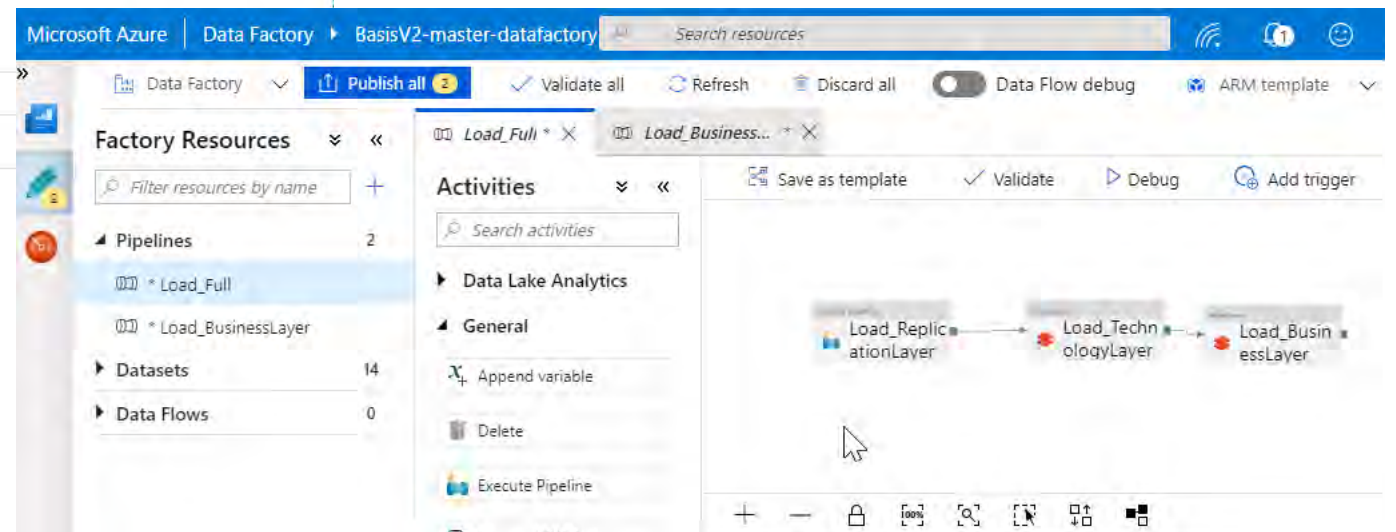
Power BI Realtime Dashboard



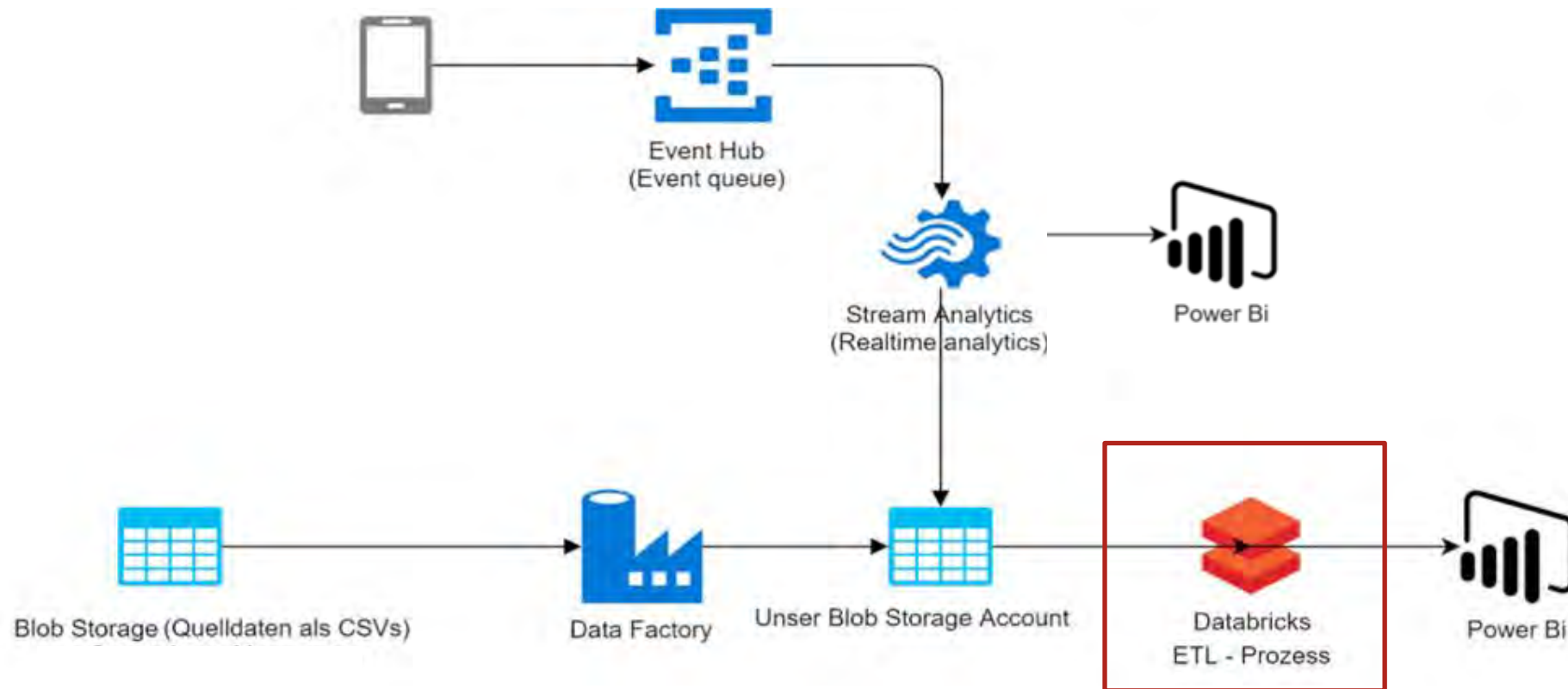
Azure Data Factory



Komplettierung Data Factory – Full Load



Praxis-Projekt MDWH Architektur



Azure Databricks



Databricks ist ein Plattform-as-a-Service (PaaS) Azure Dienst, welches auf dem Apache Spark Open Source Projekt beruht.

In Databricks lassen sich ETL / ELT Workloads flexibel in Python, Scala oder SQL abbilden. Der Vorteil von Spark bzw. Databricks liegt in der einfachen Skalierbarkeit, wodurch auch Big Data Szenarien einfach umzusetzen sind.

Azure Databricks

Microsoft Azure | Databricks

Portal j.kroehnert@oraylis.de

03. ReplicationLayer (Python)

Database: rl Path: abfss://datalake@bibasisv

```
1 %python
2
3 db = dbutils.widgets.get("database")
4 path = dbutils.widgets.get("path")
5 tbl_prefix = dbutils.widgets.get('prefix')
6
7 ## LISTING ALL FILES IN AZURE STORAGE CONTAINER
8 storage = dbutils.fs.ls(path)
9
10 ## LISTING ALL FILES IN STORAGE CONTAINER AND CREATING TABLES
11 for folder in storage:
12     foldername = folder.name
13     tablename = foldername.replace("/", ".")
14     spark.catalog.createTable(db + "." + tablename, path + tablename, source = "csv", header="True", inferSchema = "False", sep=";")
15     print("Following table was loaded: " + tablename)
```

(7) Spark Jobs

Following table was loaded: DimCustomer
Following table was loaded: DimProduct
Following table was loaded: DimPromotion
Following table was loaded: DimReseller
Following table was loaded: DimSalesTerritory
Following table was loaded: FactInternetSales
Following table was loaded: FactResellerSales

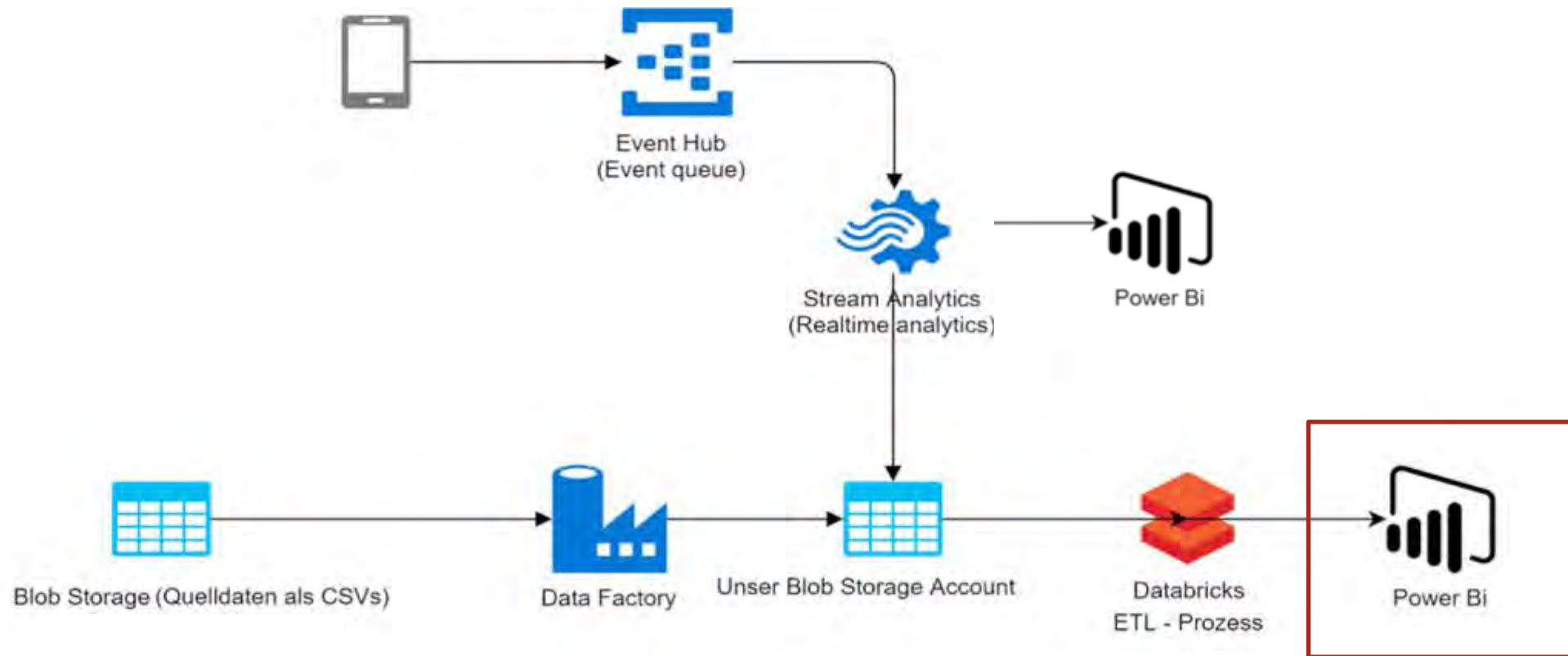
Command took 5.64 seconds -- by j.kroehnert@oraylis.de at 11/28/2022, 07:29:52 PM on DefaultCluster

```
1 %sql
2 select * from rl.DimCustomer limit 100
```

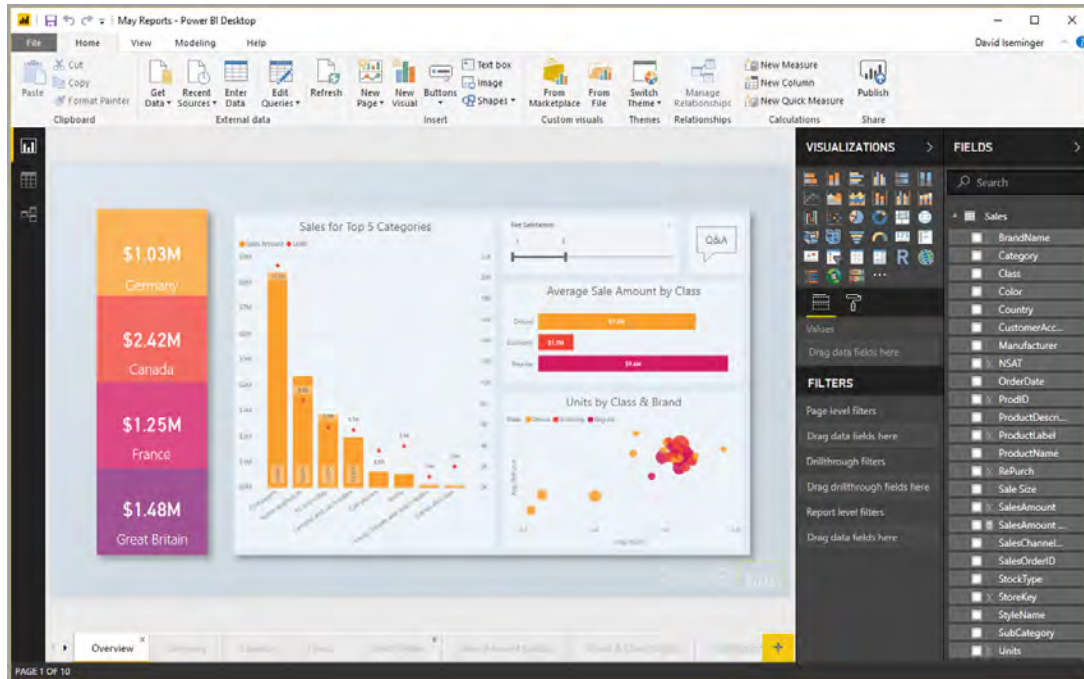
(1) Spark Jobs

	CustomerKey	CustomerID	CustomerTitle	FirstName	MiddleName	LastName	CustomerFullName	Phone	AddressLine1	AddressLine2	PostalCode	City	Country	RowsCurrent	Valid
1	1	11000	NULL	Jon	V	Yang	Yang, Jon V	1 (11) 500 555-0162	3761 N. 14th St	NULL	4700	Rockhampton	Australia	NULL	8/30/
2	2	11001	NULL	Eugene	L	Huang	Huang, Eugene L	1 (11) 500 555-0110	2243 W St.	NULL	3198	Seaford	Australia	NULL	8/30/
3	3	11002	NULL	Ruben	NULL	Torres	Torres, Ruben	1 (11) 500 555-0184	5844 Linden Land	NULL	7001	Hobart	Australia	NULL	8/30/
4	4	11003	NULL	Christy	NULL	Zhu	Zhu, Christy	1 (11) 500 555-0162	1825 Village Pl.	NULL	2113	North Ryde	Australia	NULL	8/30/

Praxis-Projekt MDWH Architektur



Desktop Power BI



Daten abrufen

1

databricks

Alle

Azure Databricks

Azure

Azure Databricks

2

Server Hostname

Beispiel: example.azuredatabricks.net

HTTP Path

Beispiel: sql/protocolv1/o/1814582234607533/7508-187377-agent704

Advanced Options (optional):

Datenkonnektivitätsmodus

Importieren

DirectQuery

Navigator

3

Anzeigeoptionen

adb-6000646038357221.1.azuredatabricks.net:

SPARK [4]

bl [9]

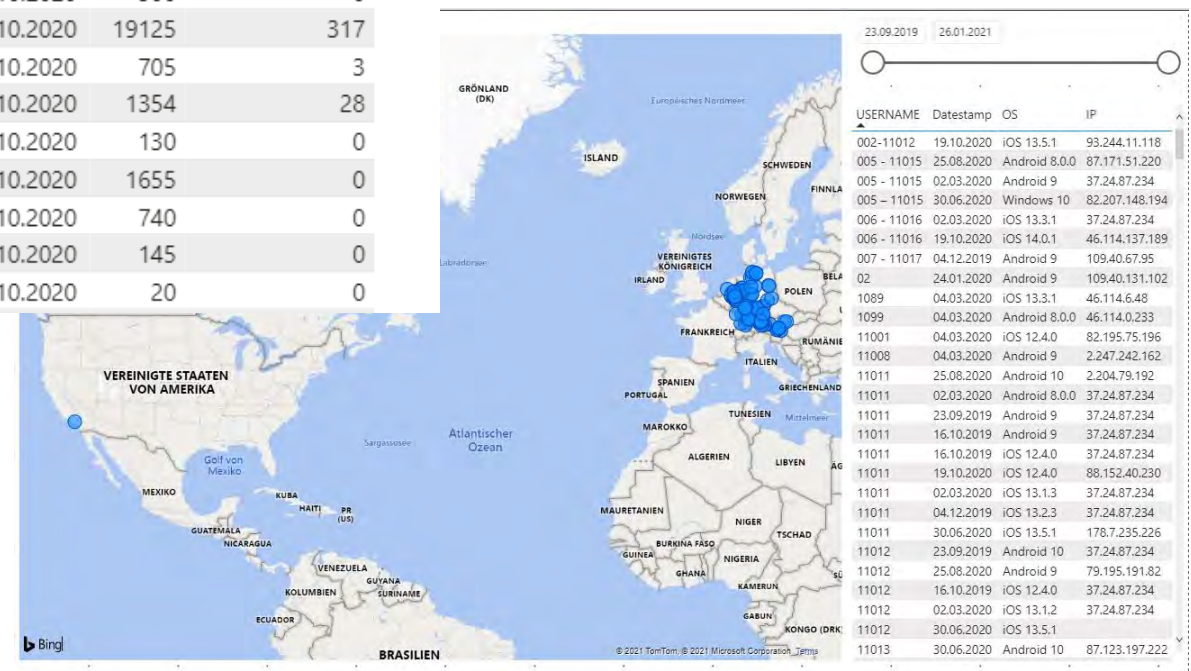
dimcustomer
dimdate
dimproduct
dimpromotion
dimreseller
dimesalesterritory
factinternetsales
factresellersales
factshake

Power BI ist ein visuelles Business Analytics Werkzeug mit dem Daten aufbereitet und visualisiert werden können. Power BI läuft auf der sog. Tabular Cube Engine – diese ist ein Teil von Power BI welche es ermöglicht Daten blitzschnell inMemory zu verarbeiten.

Power BI Portal

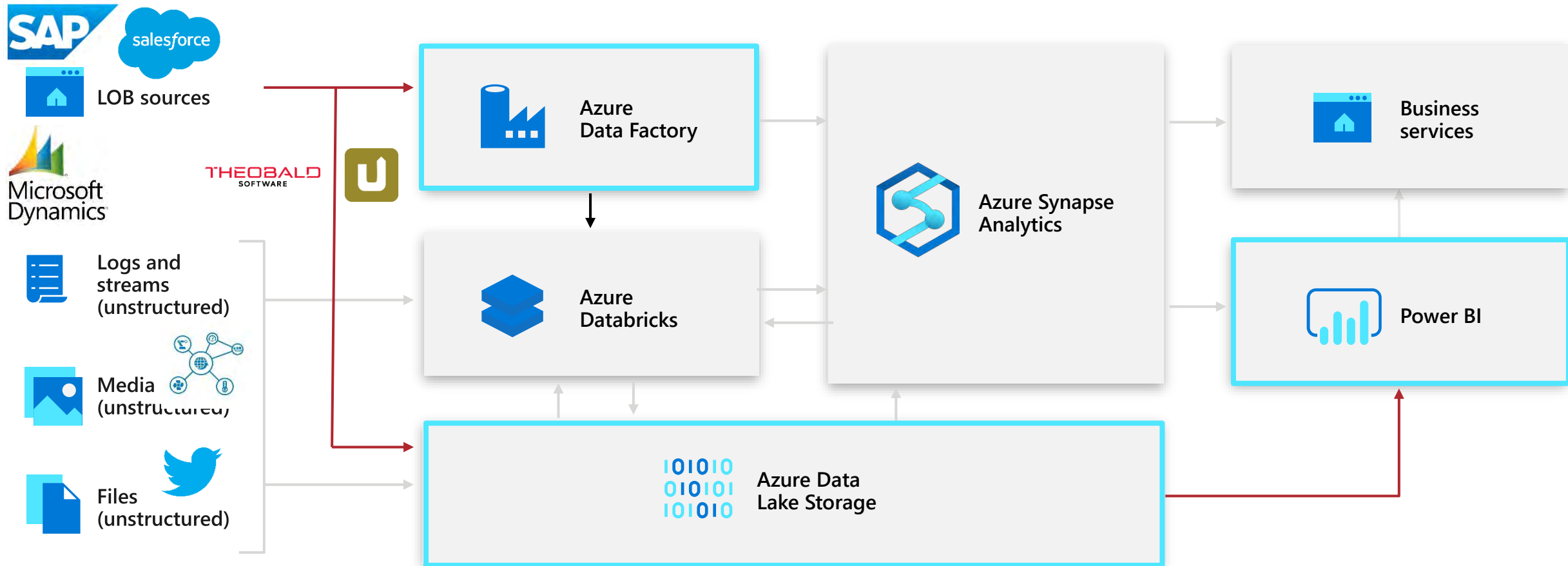
Power BI Report(s)

USERNAME	FirstName	LastName	city	IP	LATITUDE	LONGITUDE	Datestamp	VOTE	AVERAGESHAKE
11111	Eduardo	King	Beaverton	5.147.1.70	51.22441968092483	6.915386548348868	26.01.2021	1616	4
11017	Terrence	Andersen	Burien	109.40.64.82	51.2123875	6.7647575	21.10.2020	580	0
11017	Terrence	Andersen	Burien	109.40.64.82	51.2123875	6.7647575	20.10.2020	900	0
11011	Henry	Lopez	Burlingame	88.152.40.230	51.224503159724115	6.915563118094766	19.10.2020	19125	317
11014	Marcus	Jenkins	Milwaukie	37.24.87.234	51.2731176	6.7678508	19.10.2020	705	3
11014	Marcus	Jenkins	Milwaukie	37.24.87.234	51.273119	6.7678582	19.10.2020	1354	28
11015	Jose	Gonzales	Haney				19.10.2020	130	0
11015	Jose	Gonzales	Haney		51.254505826525275	6.807730569342475	19.10.2020	1655	0
11015	Jose	Gonzales	Haney		51.254511359031554	6.80772766036329	19.10.2020	740	0
11015	Jose	Gonzales	Haney		51.254513544356634	6.8078031769703795	19.10.2020	145	0
11017	Terrence	Andersen	Burien	109.40.64.82	51.2123123	6.7646982	19.10.2020	20	0



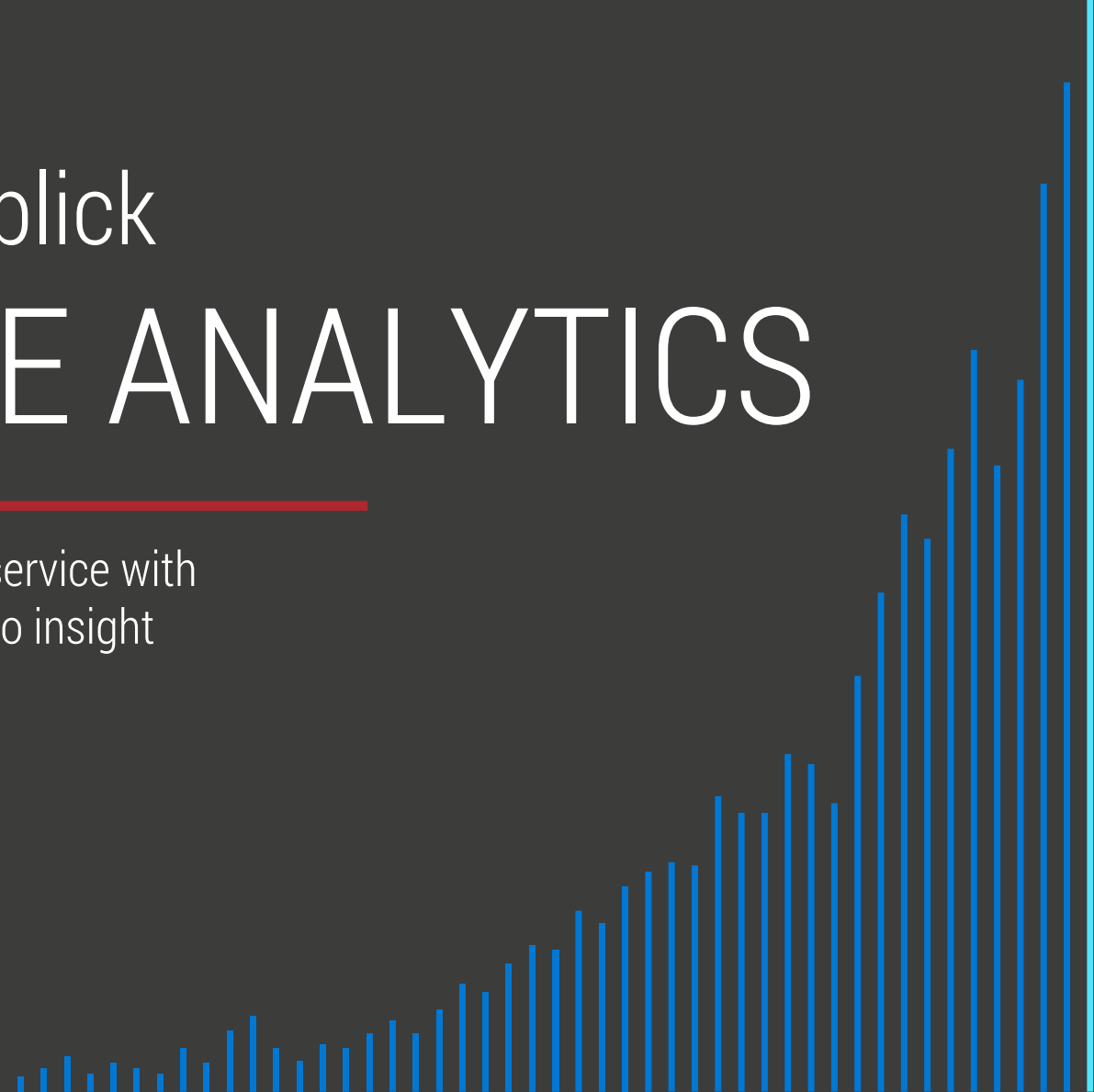
Importierung der Business Layer Tabellen aus Databricks nach Power BI und Erstellung von Reports

Azure Synapse



Ein Überblick AZURE SYNAPSE ANALYTICS

Limitless analytics service with
unmatched time to insight





Einführung in Azure Synapse Analytics

Azure Synapse ist ein **skalierbarer** Analysedienst, der **Data Warehousing** für Unternehmen mit **Big Data-Analysen** **kombiniert**.

Einfach ausgedrückt: **Azure Synapse ist das weiterentwickelte Azure SQL Data Warehouse**

Microsoft hat das branchenführende Data Warehouse auf ein völlig neues Leistungsniveau gehoben.

Cloud-scale analytics



Modern data warehousing

"We want to analyze data coming from multiple sources and in varied formats"



Advanced analytics

"We want to leverage the analytics platform for advanced fraud detection"



Real-time analytics

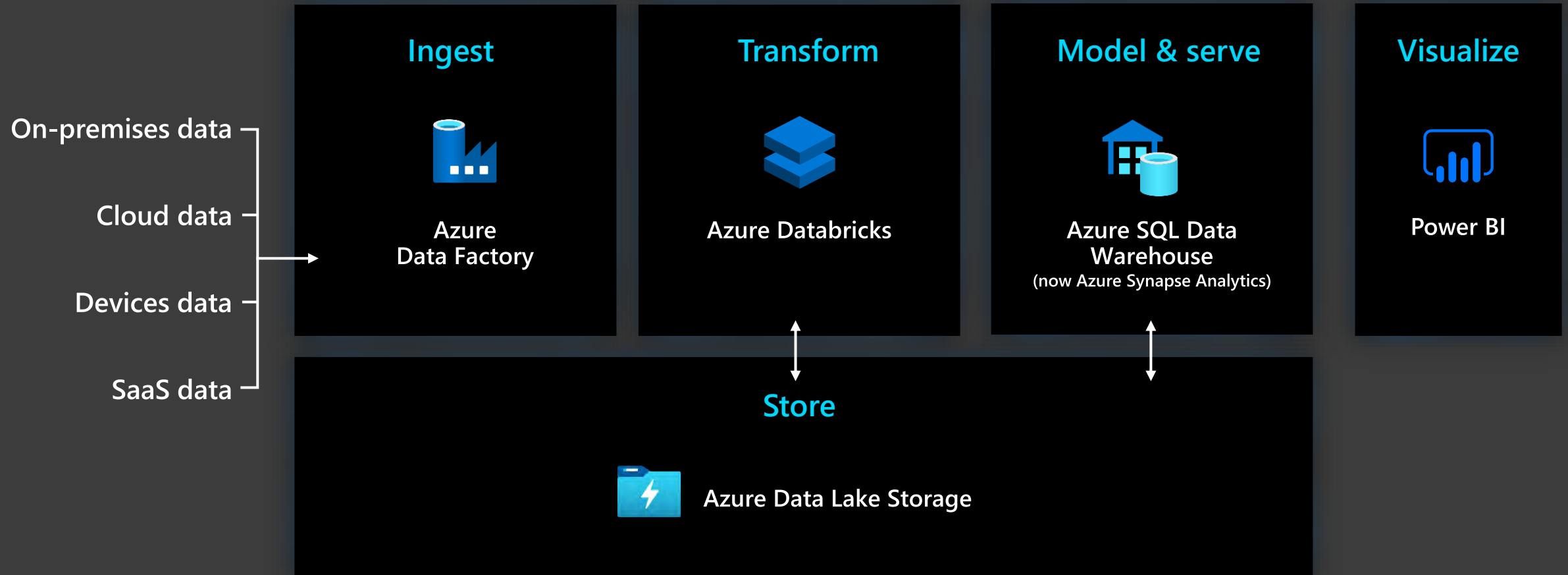
"We're trying to get insights from our devices in real time"

Azure Synapse Analytics

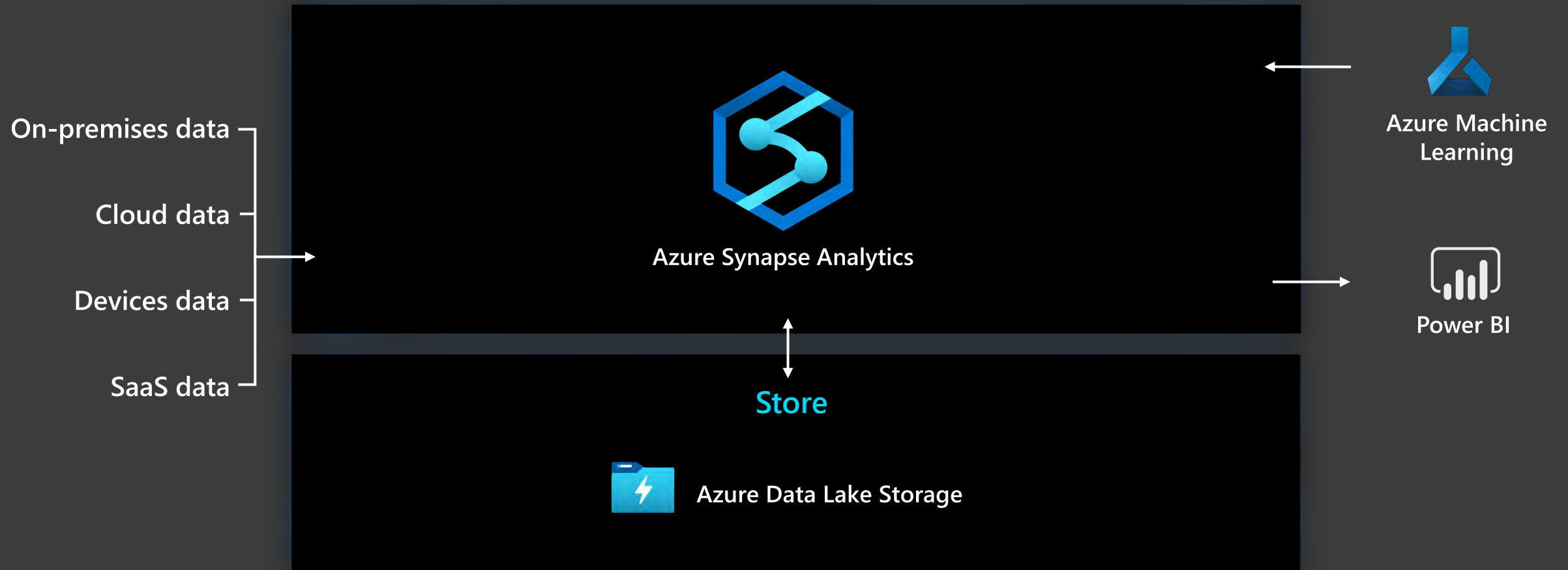
Die

VISION DAHINTER

Modern Data Warehouse



Azure Synapse Analytics



Was ist
DRIN?

Azure Synapse Analytics



Synapse SQL



Apache Spark
for Synapse



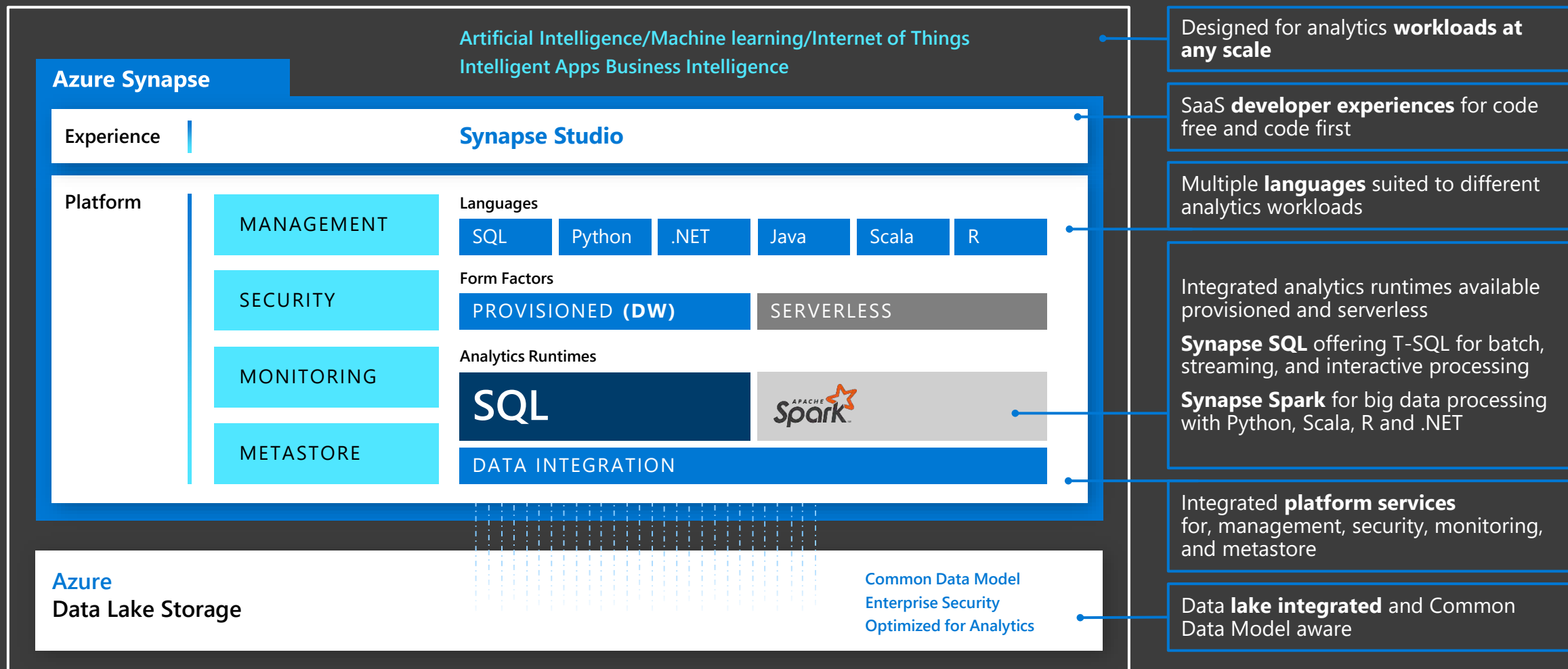
Synapse Pipelines



Synapse Studio

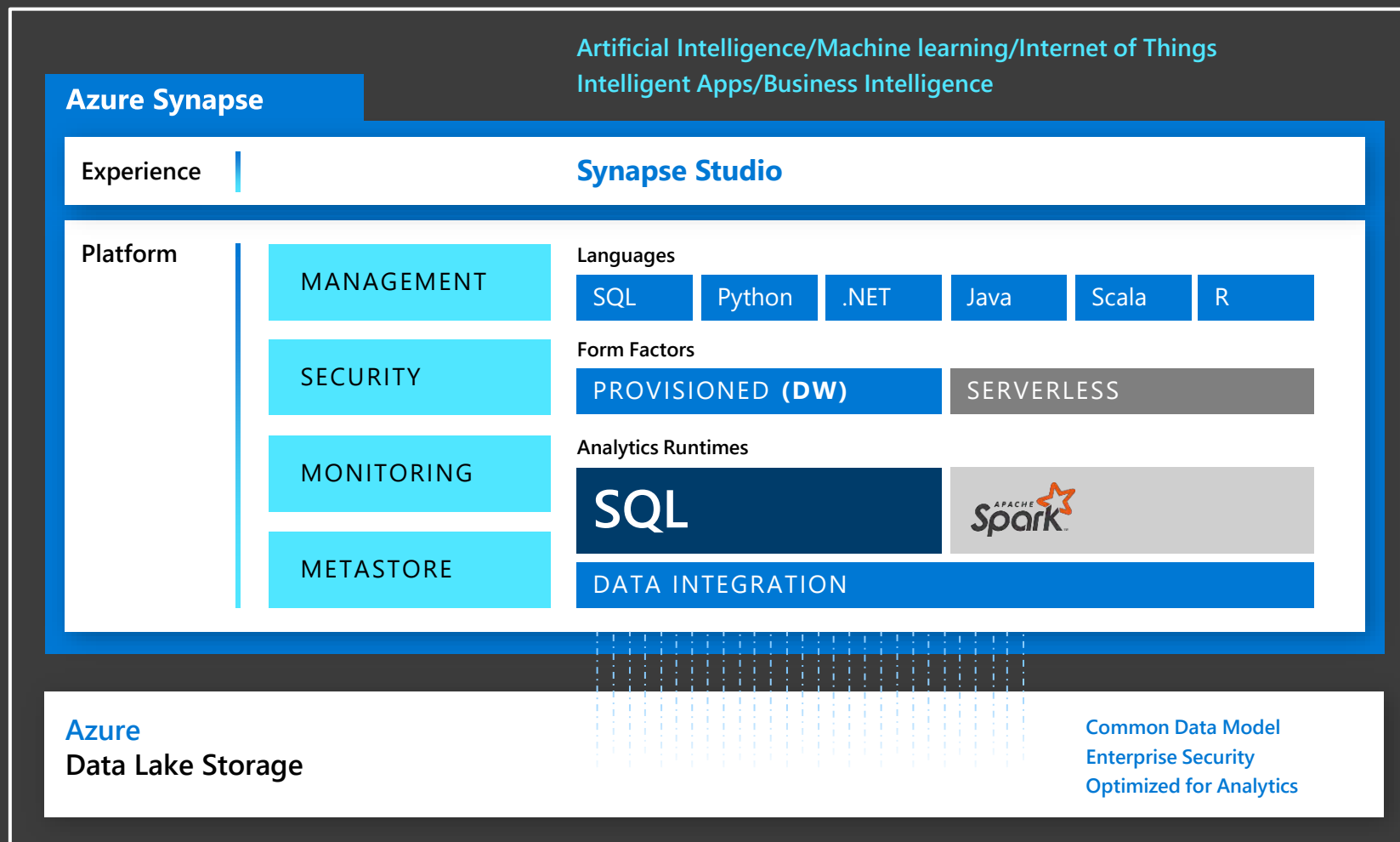
Azure Synapse Analytics

Integrated analytics platform for AI, BI, and continuous intelligence



Azure Synapse Analytics

Integrated analytics platform for AI, BI, and continuous intelligence



Connected Services

Azure Purview (Data Catalog)
Azure Data Lake Storage
Azure Data Share
Azure Databricks
Azure HDInsight
Azure Machine Learning
Power BI

3rd Party Integration

Azure Synapse Analytics



Synapse SQL



Apache Spark
for Synapse



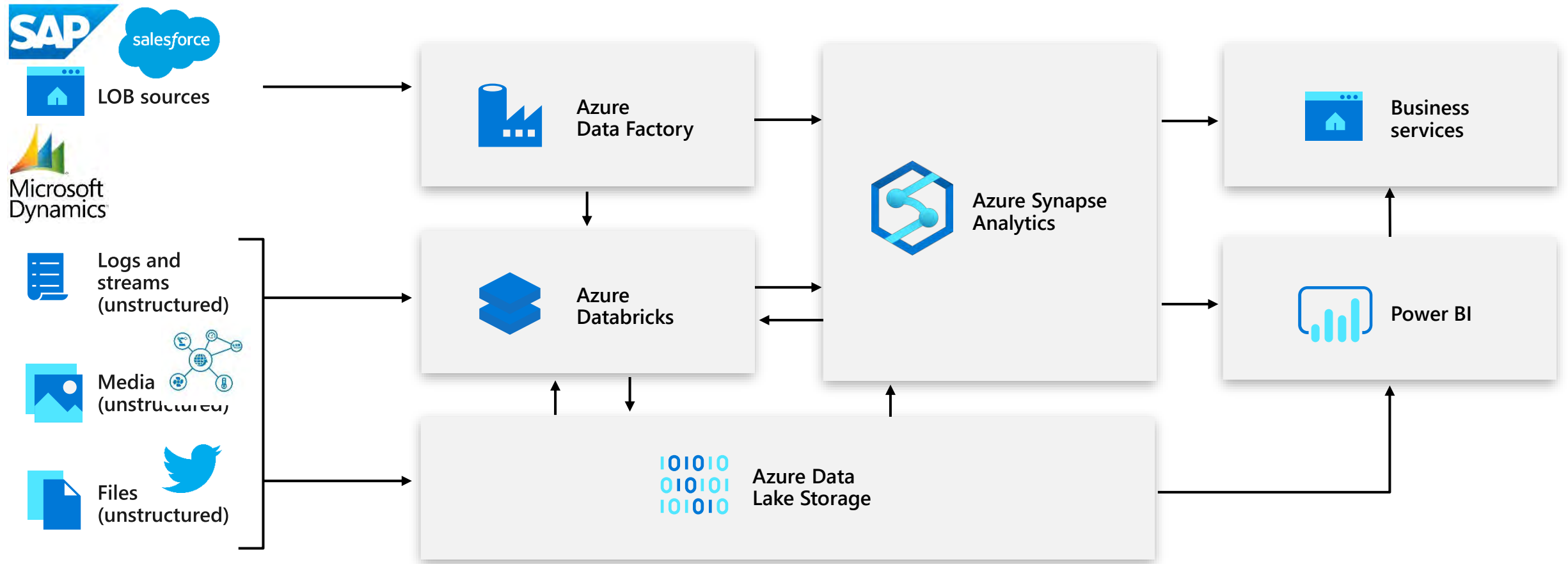
Synapse Pipelines



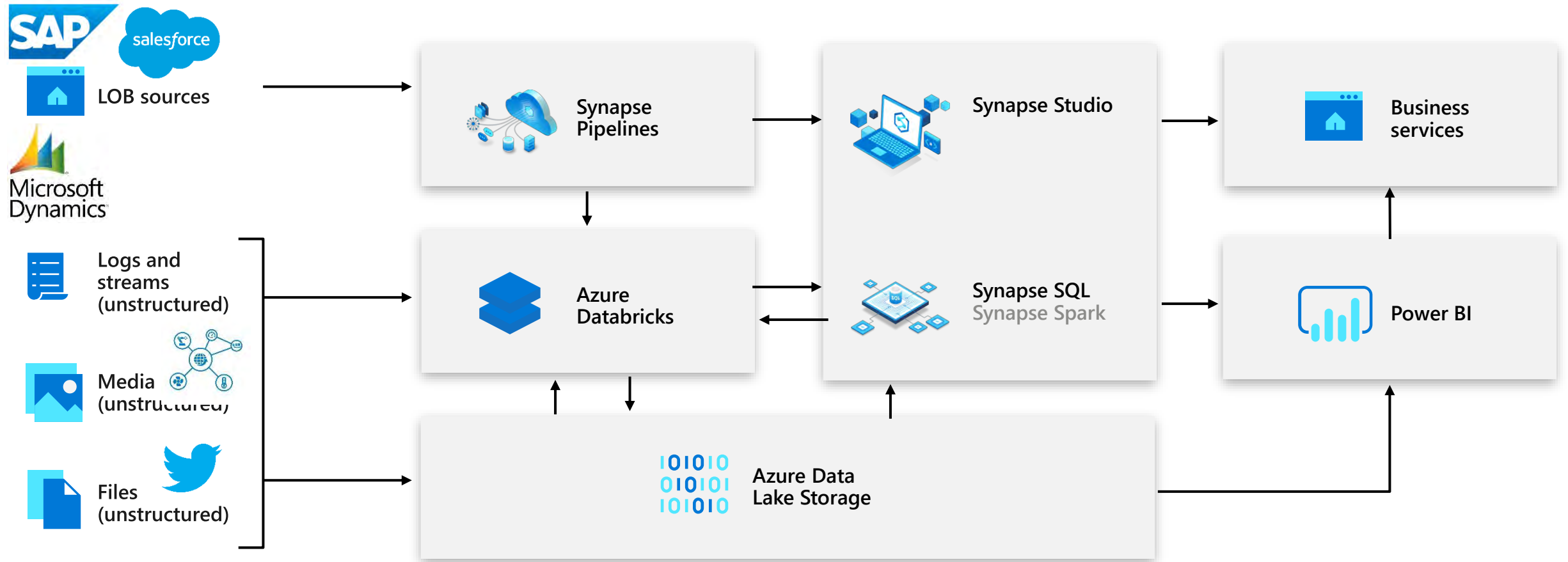
Synapse Studio

Welchen Impact hat das auf mögliche
MDWH-Architekturen

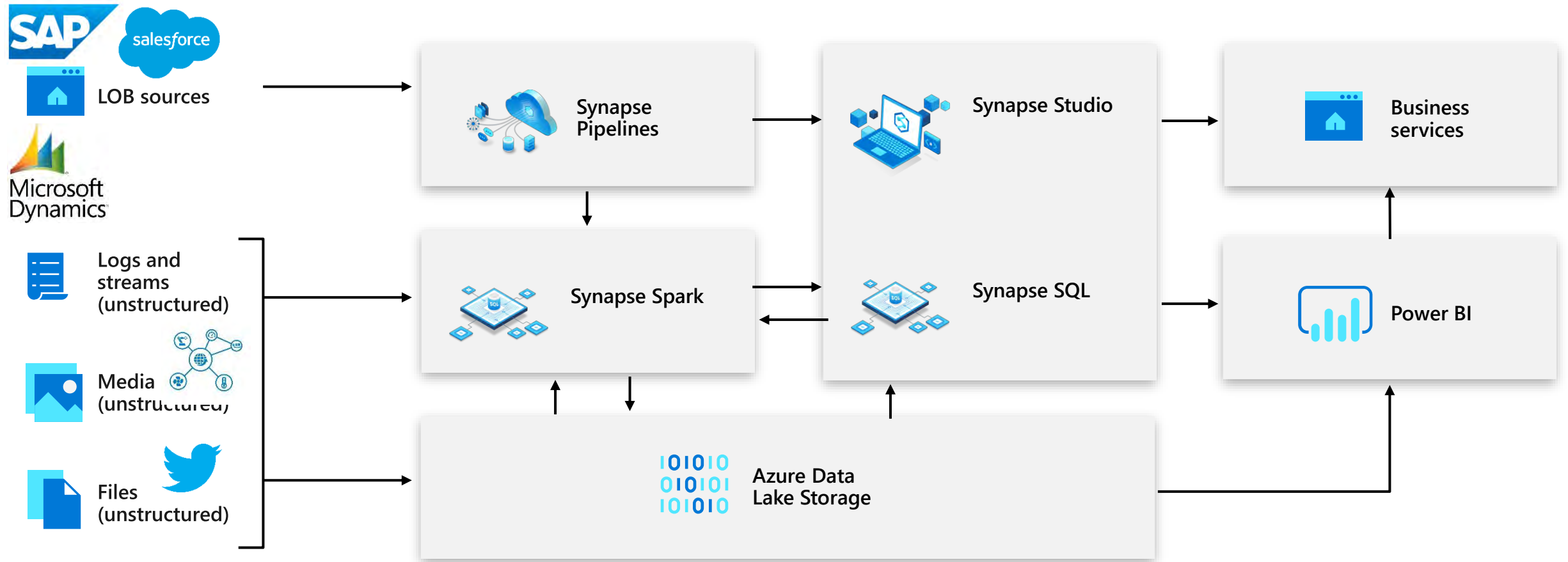
AZURE MODERN DATA WAREHOUSE 1.0



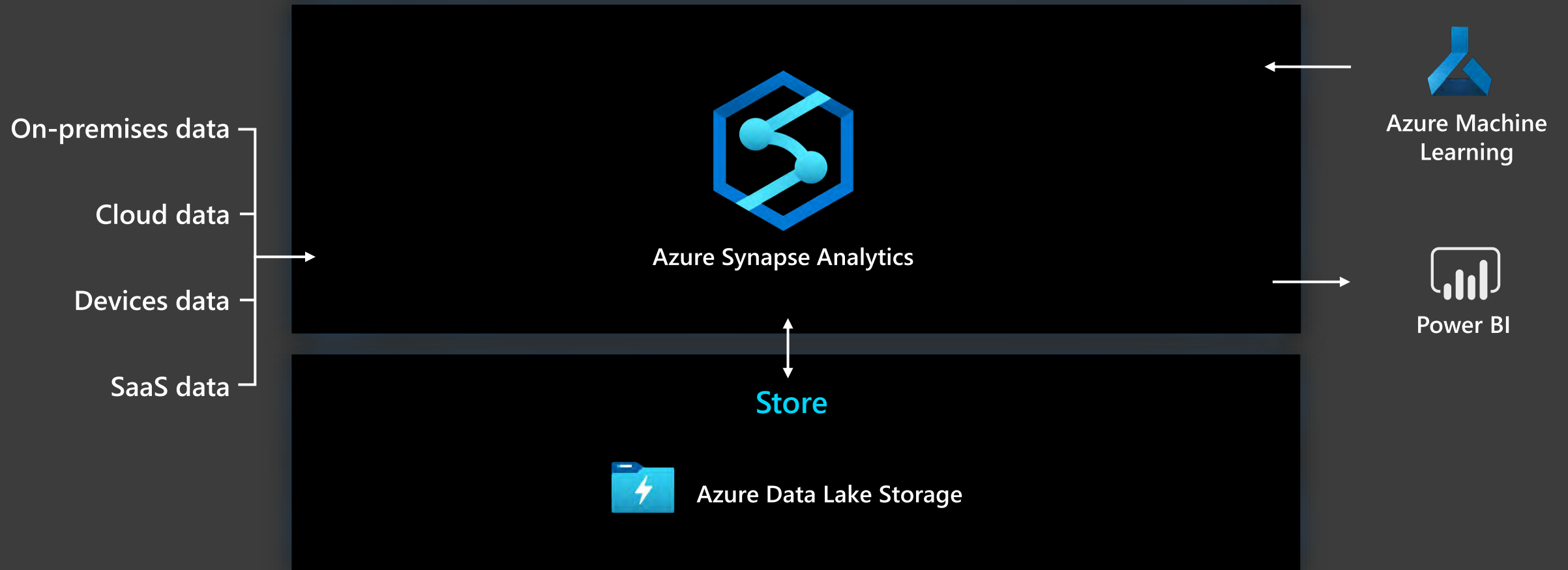
AZURE MODERN DATA WAREHOUSE 1.1



AZURE MODERN DATA WAREHOUSE x.x



Azure Synapse Analytics



MICROSOFT IST NICHT ALLEINE – TESTEN TESTEN

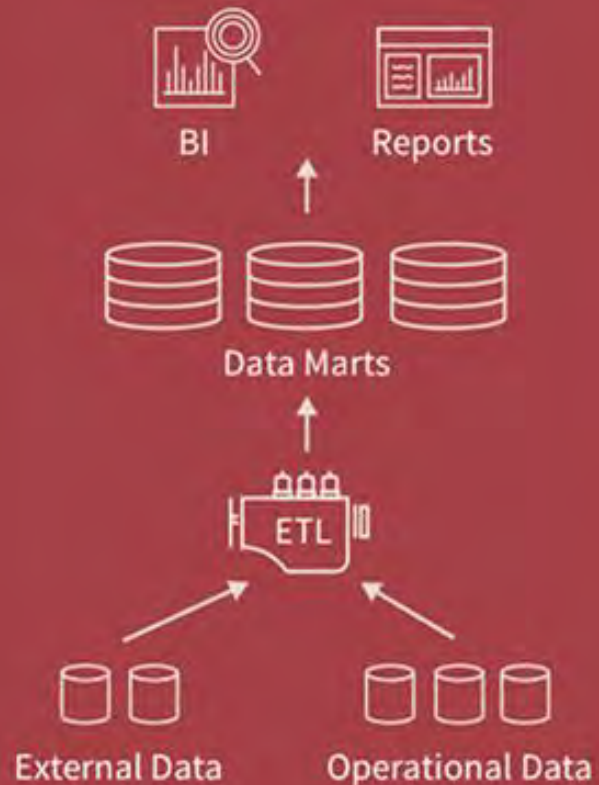
Synapse serverless	Synapse dedicated	Databricks	Dremio
Power BI Composite Model	Power BI Composite Model	Power BI Composite Model	Power BI Composite Model
Synapse serverless DWUs ? ...	Synapse dedicated DWUs ? Distribution ? ...	Databricks # Nodes ? Nodesizes ? ..	Dremio # Cluster size ? Reflections ?
Datalake Store Driver ? Partitioning ? Fileformats ? .. ?		Datalake Store Partitioning ? Fileformats ? .. ?	Datalake Store Partitioning ? Fileformats ? .. ?

Wie verändern sich die
DATEN-MODELLE?

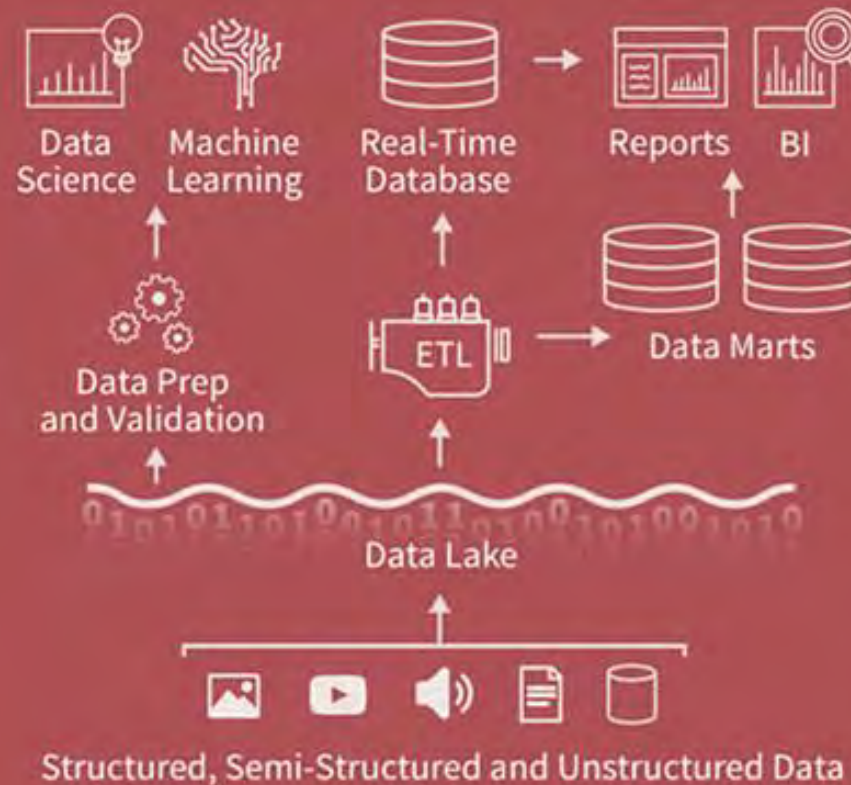
Das Data Warehouse...

GESTERN, HEUTE & MORGEN

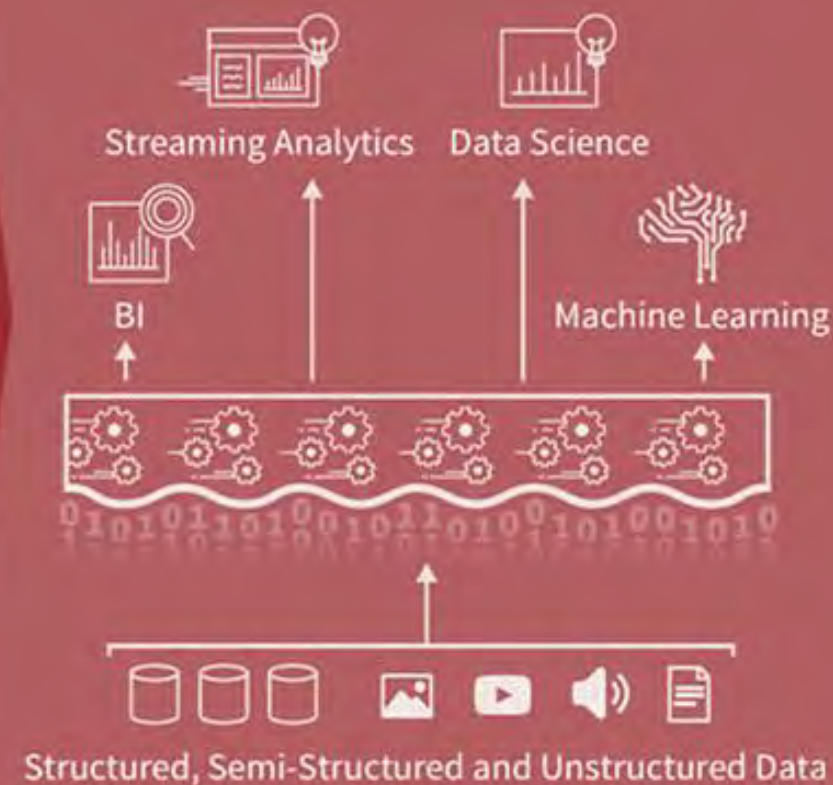
KLASSISCH



DATA LAKE



VIRTUALISIERT



UNSERE KONTAKTINFOS



Daniel Esser
Senior Consultant



Jens Kröhnert
Principal Consultant



Q&A
