

# Azure Data Factory – Data Management für die Cloud

## Stefan Kirner

PASS Treffen Regionalgruppe Bayern  
10.03.2016, Microsoft Unterschleißheim

# Pre-Talk Teaser – der heiße Scheiss!

```
|select @@version
```

100 %

Results

```
-----  
Microsoft SQL Server (Preview) - 13.0.8000.6 (X64)  
Feb 24 2016 22:03:46 2015.0130.8000.06  
Copyright (c) Microsoft Corporation  
on Linux (Ubuntu 15.10)
```

```
(1 row(s) affected)
```

I

# Speaker Bio: Stefan Kirner

- Teamleiter BI Solutions bei der inovex GmbH
- Erfahrung mit Microsoft BI Stack seit SQL Server 2000
- Entwicklung von Data Management Lösungen in der Cloud
- Microsoft Certified Systems Expert (MCSE)
- Leitet seit 2006 die PASS Regionalgruppe Karlsruhe
- Unterstützt Microsoft Deutschland als P-TSP im Presales Bereich
- Kontakt: [stefan.kirner@inovex.de](mailto:stefan.kirner@inovex.de)



# Data Factory ist einfach...

Mein Sohn Leo beim Bau seiner ersten  
Data Factory!



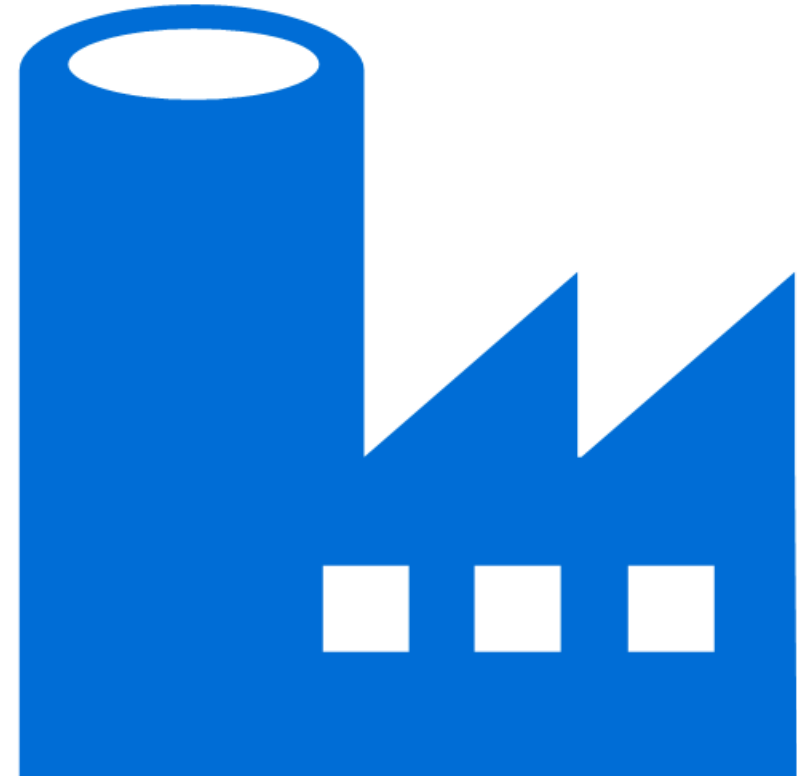
# Es funktioniert!



# ...die Wirklichkeit ist etwas komplizierter...

## Die Agenda:

- Warum Azure Data Factory?
- Beispielszenarien
- Elementare Begriffe
- Mit Data Factories arbeiten
- Live Demo
- Kosten
- Vertiefungen

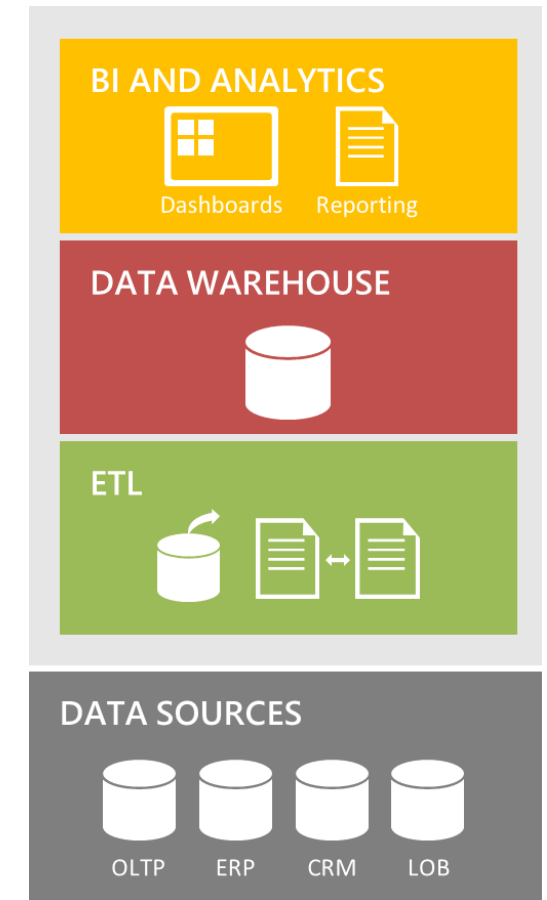


# Warum Azure Data Factory?

# Das “traditionelle” Data Warehouse

“...data warehousing has reached the most significant tipping point since its inception. The biggest, possibly most elaborate data management system in IT is changing.”

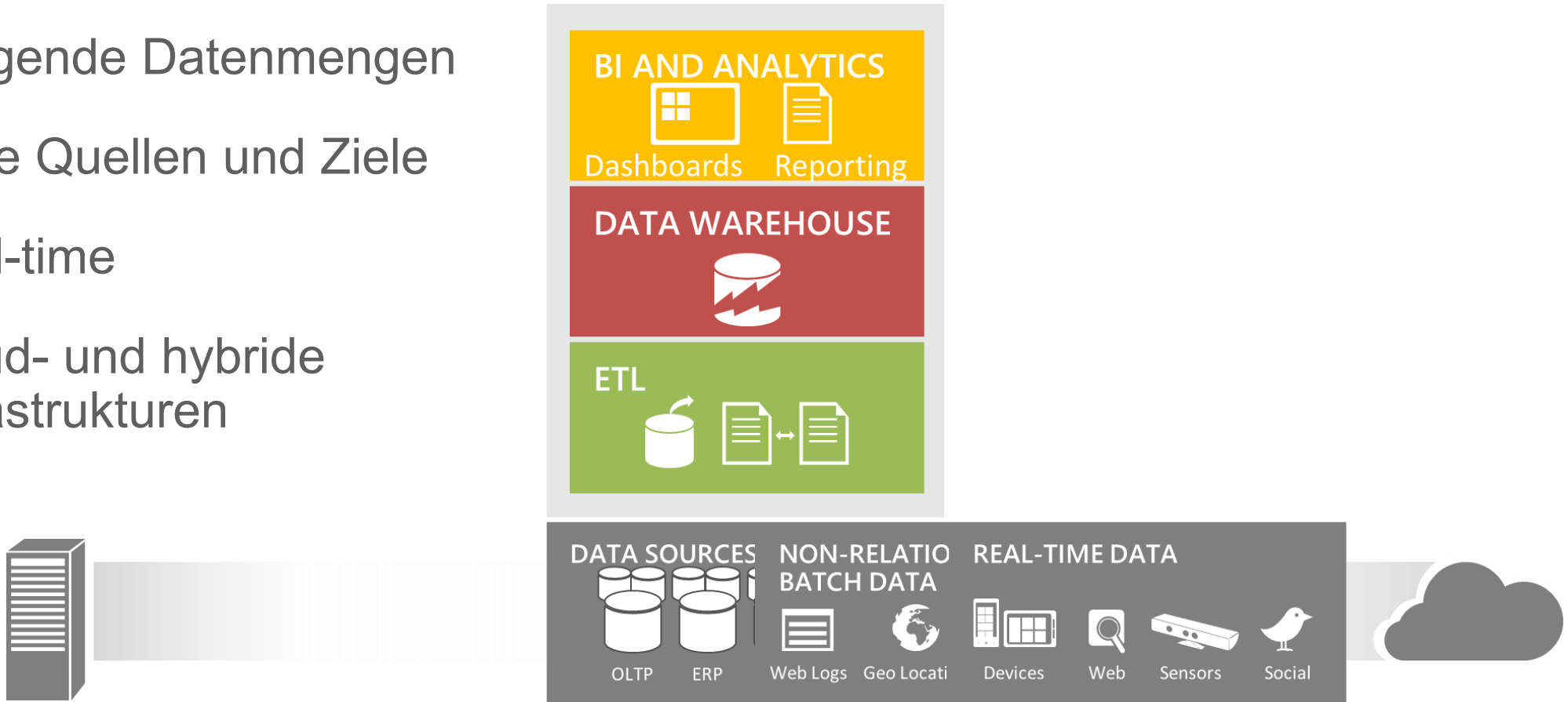
GARTNER, “THE STATE OF DATA WAREHOUSING IN 2012”



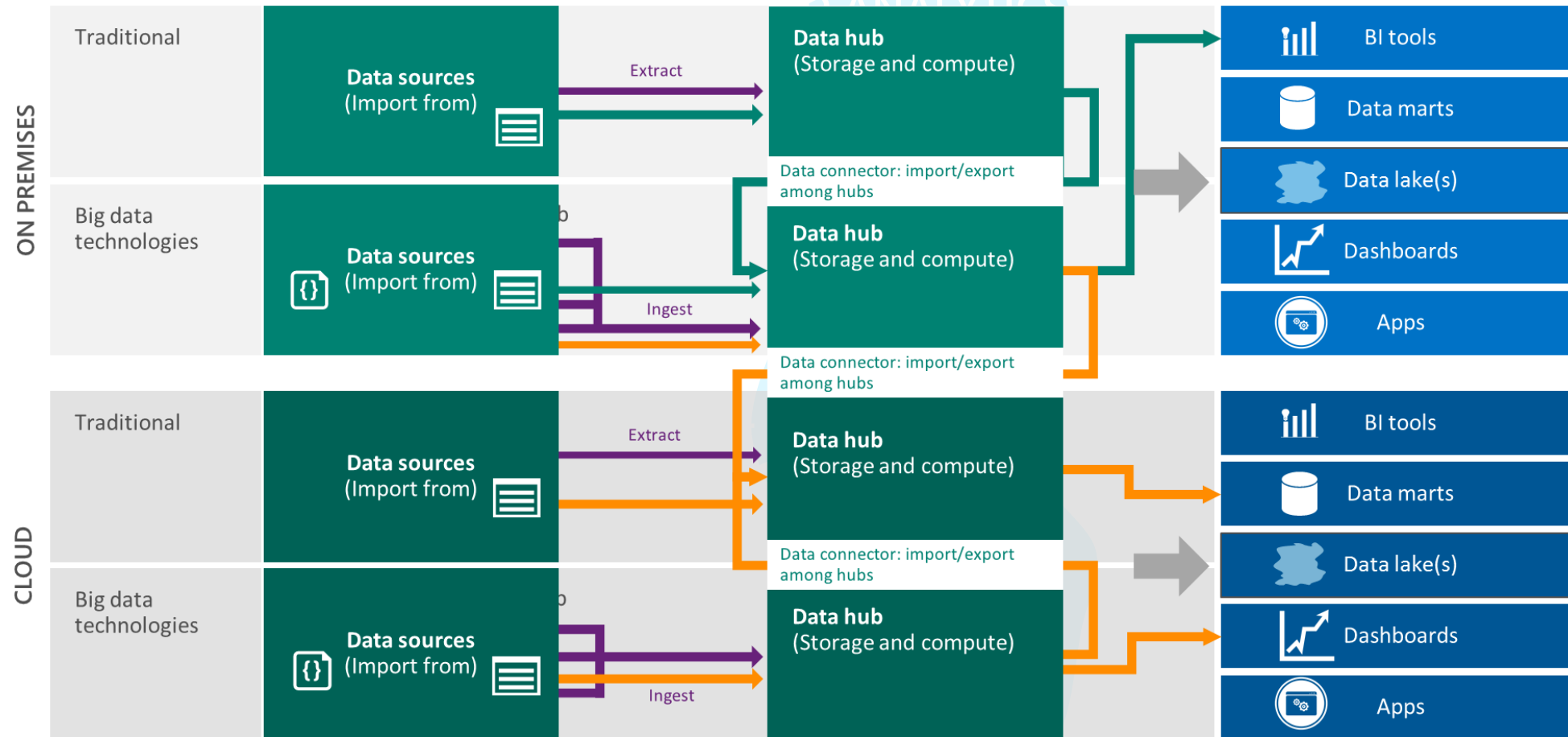


# Data Warehousing verändert sich

- 1 Steigende Datenmengen
- 2 Neue Quellen und Ziele
- 3 Real-time
- 4 Cloud- und hybride Infrastrukturen

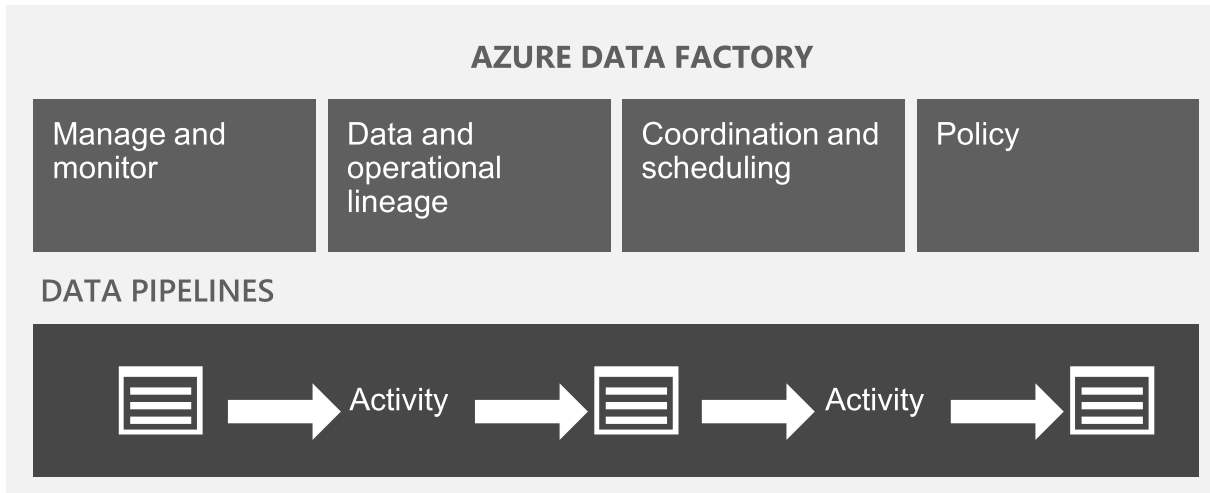


# Analytische Workflows werden immer komplexer



*Es folgt ein kurzer Reklame-Block ;-)*

# Announcing Azure Data Factory



## Compose and orchestrate

- Connect to data of diverse shapes and location
- Orchestrate processing to produce trusted data

## Manage from single pane of glass

- Manage a network of data pipelines
- See lineage and impact analysis

## Set data production policy

- Retry, concurrency, late data handling

## Identify and debug errors

- Automatic dataset health alerts
- Troubleshoot complex pipelines

# Data Factory ist Teil der Cortana Analytics Suite



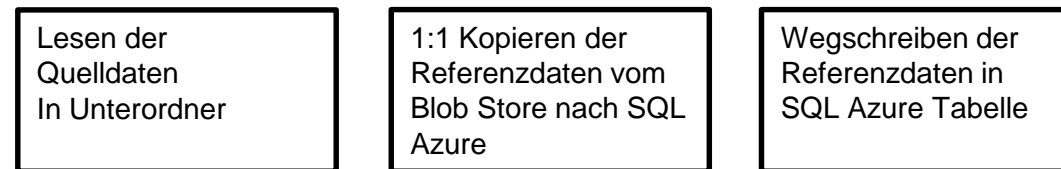
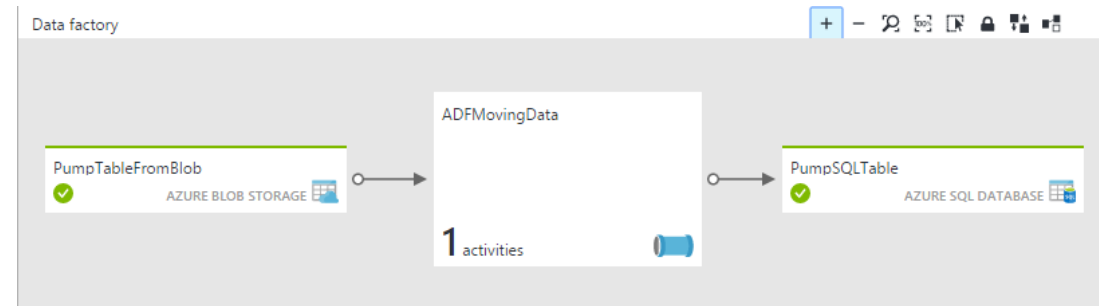
# Azure Data Factory ist nicht...

*...SSIS in der Cloud!*

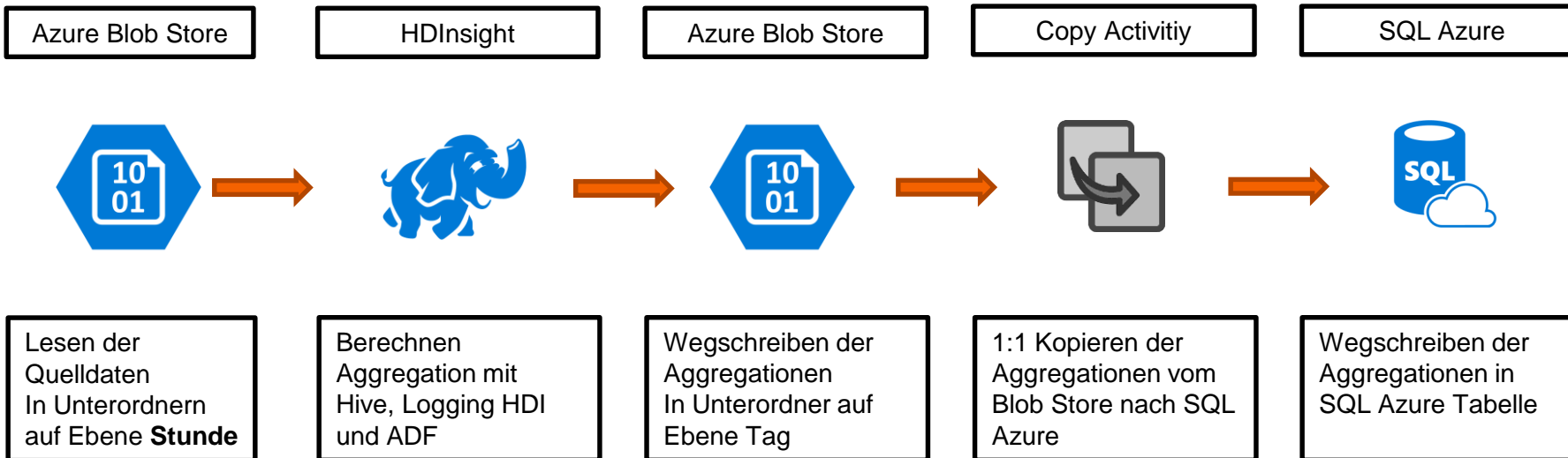
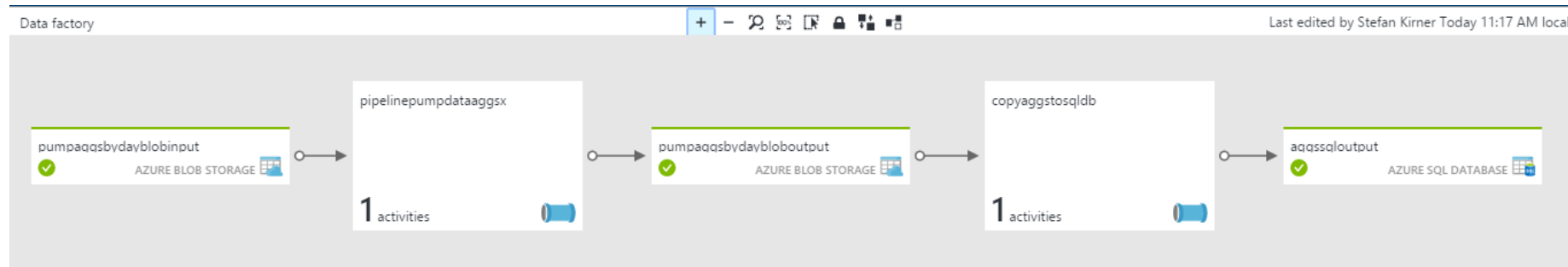
	SSIS	Data Factory
Einsatzziel	Traditionelles ETL-Tool	Orchestrierungsdienst
Skalierbarkeit	Für DWH ausreichend	Für Big Data gedacht
„Reichweite“	Firmennetzwerk	Global in Azure
Art der Software	On Prem Software	Managed Service
Produktzyklen	Lang: Jahre	Kurz: Wochen

# Beispielszenarien

# Beispiel: Laden von Referenzdaten in DB



# Beispiel: Daten-Aggregation mit Hadoop

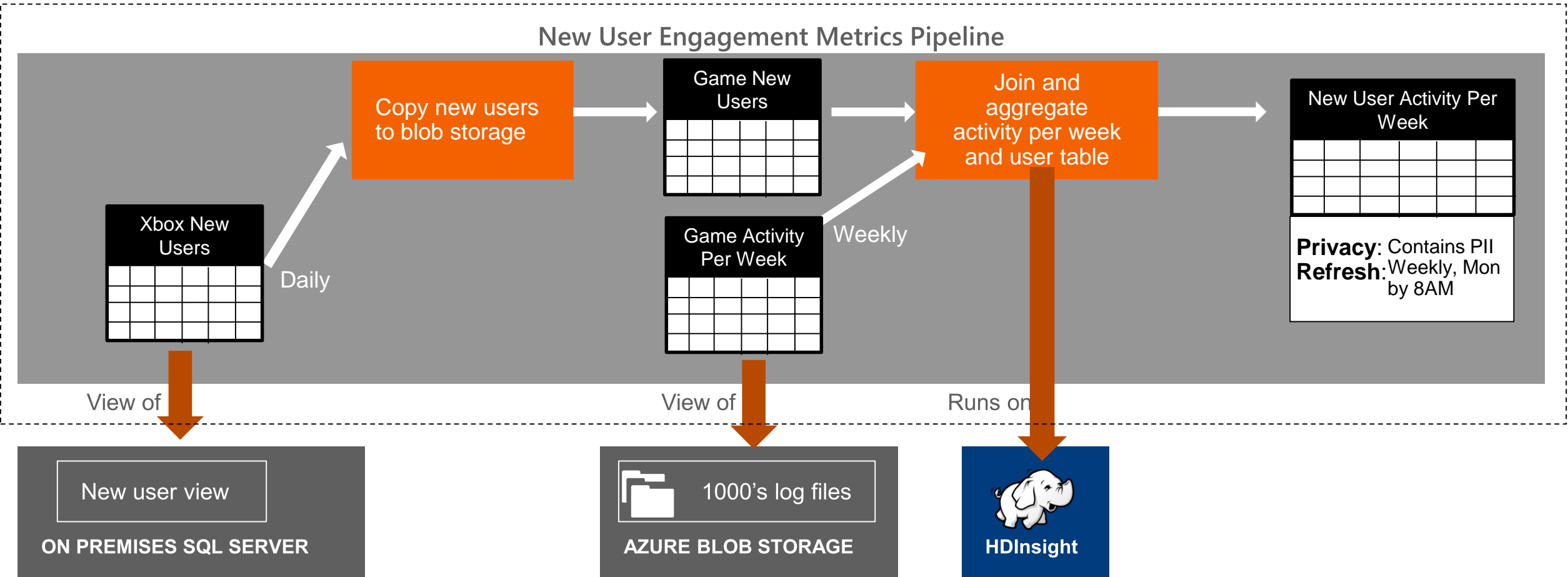




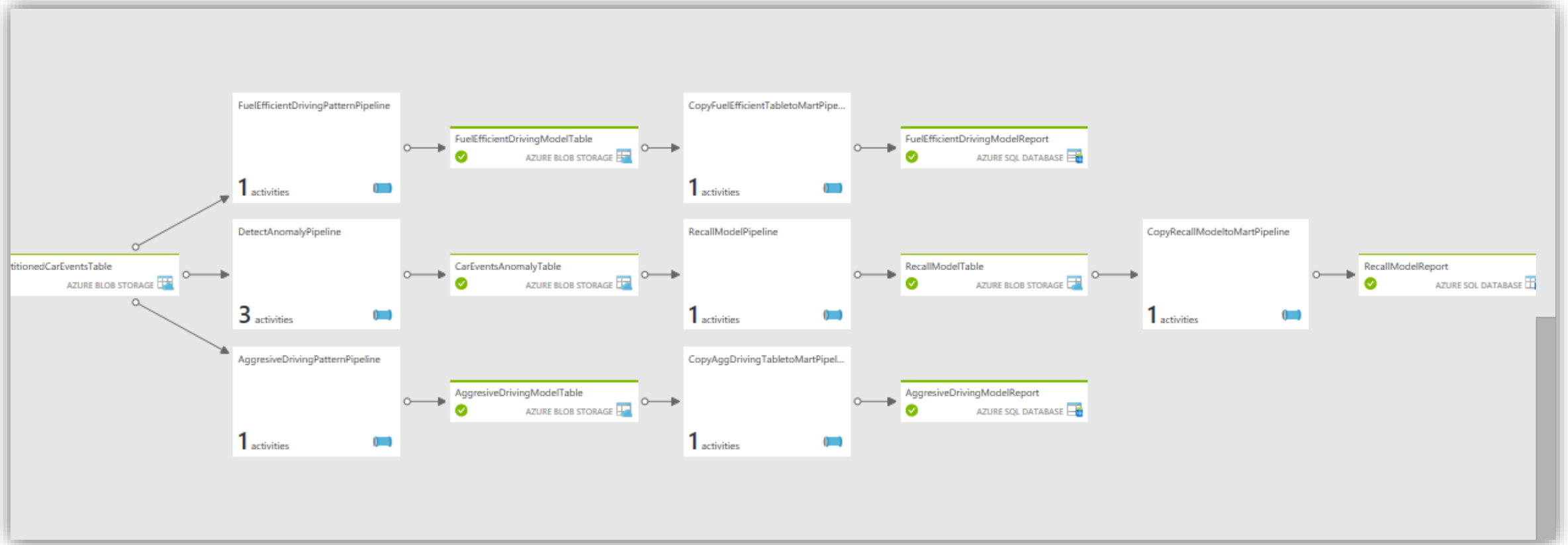
# Beispiel: Kunden-Profilung & Spiel-Analyse

## AZURE DATA FACTORY

### New User Engagement Metrics Pipeline



# Beispiel: Mehr Komplexität - Connected Cars Demo

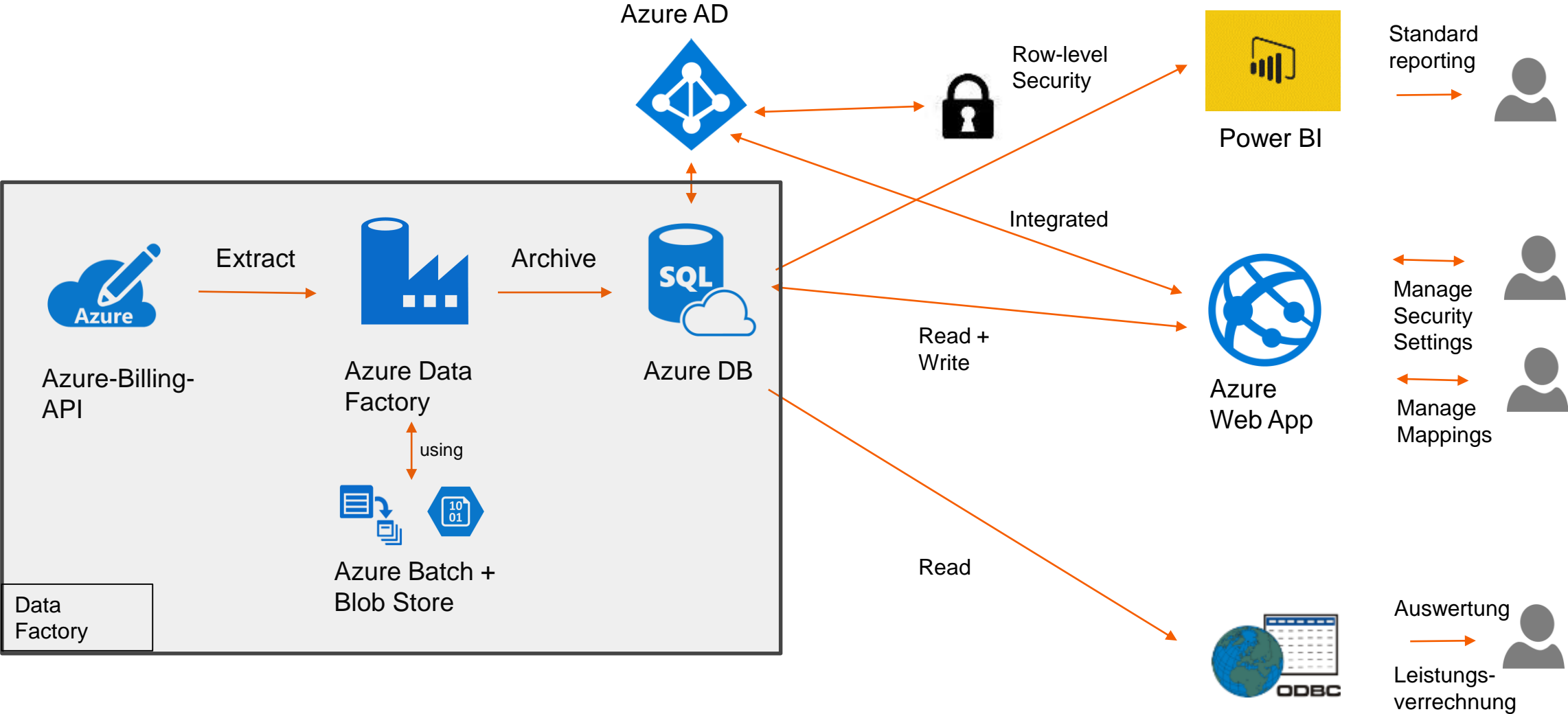


# Beispiel Kundenprojekt: Azure Consumption Reporting

- Auswertung Azure Consumption nach den unterschiedlichen Hierarchiestufen die im EA Azure Portal vorgesehen sind: Enrollment Level, Department, Account und Subscriptions
- Schnittstelle ist ein Webservice
- Datengrundlage sind Azure-Abrechnungsdaten
- Security Konzept auf Datenebene
- Pflege der Berechtigungen und Zuordnungen über Web-Apps

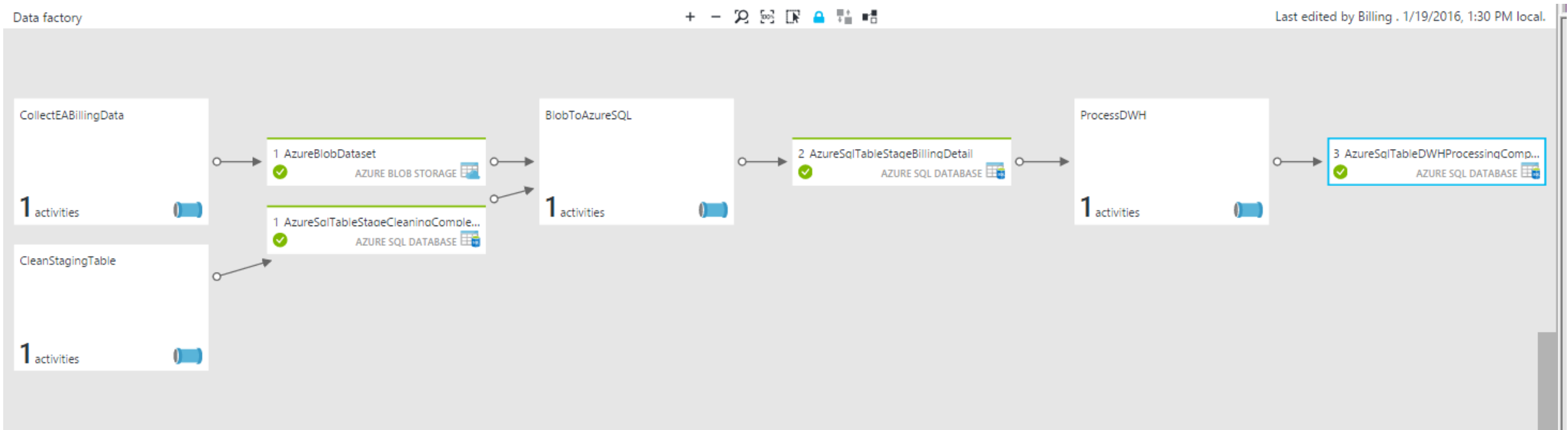
**Non-Functional Requirement: Nur PaaS Dienste verwenden um Betriebsaufwände auf Kundenseite zu vermeiden**

# Architektur



# Zooming to Data Factory part

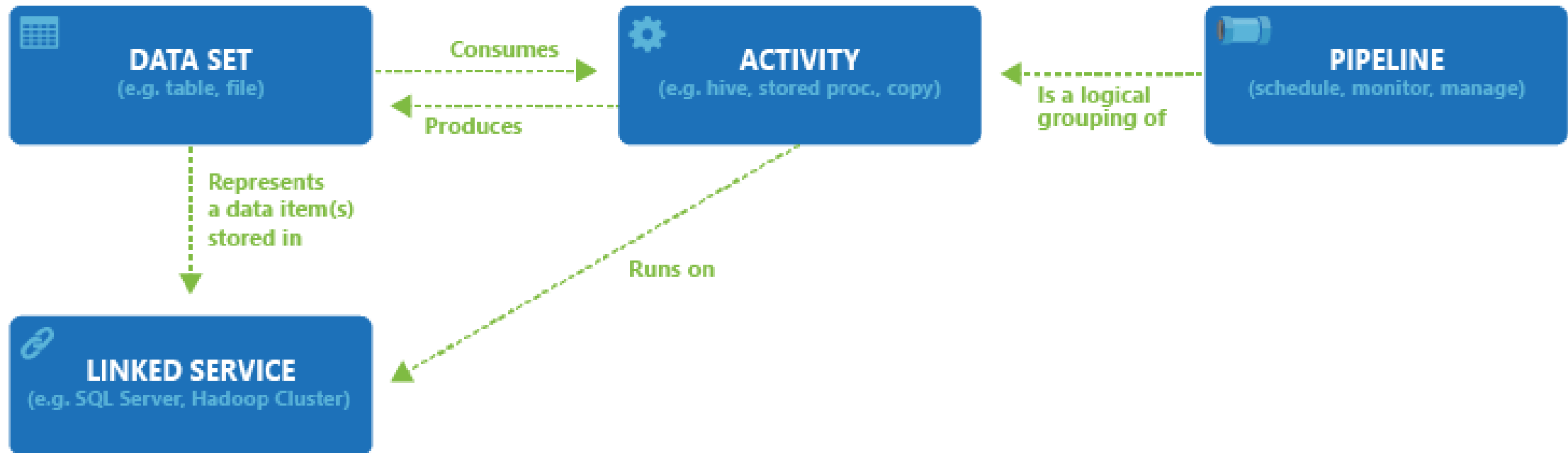
- Sourcing 1x täglich aus Web-Schnittstelle über eigenentwickelte custom .net Activity
- Nutzen von Azure Batch als Compute Knoten und Blob Store als Zwischenspeicher
- Laden nach Azure SQL für Staging und DWH Layer
- Alerting bei Fehlern



# Elementare Begriffe

- Linked Service
  - Data Store – Speichern *und Verarbeiten* von Daten (SQL, Storage,...)
  - Data Gateway – Verbindung zu On Prem-Daten (On Premises ORACLE,...)
  - Compute – Verarbeitung von Daten (Batch, HDInsight, ML,...)
- Data Set
  - Datenbeschreibung (Tabelle, Datei,...)
- Activity
  - Verarbeitet ein Dataset und erzeugt ein neues Dataset (Copy, Hive, Stored Procedure,...)
  - Entweder Data Movement oder Data Transformation
- Pipeline
  - Gruppierung von Aktivitäten, die eine bestimmte Aufgabe lösen
  - Deployment- & Management-Einheit für Aktivitäten

# Wie hängen die Bausteine zusammen?





# (Einige) Aktuell vorhandene Bausteine

## – Compute Linked Services

- HDInsight
- Azure Batch
- Azure ML
- Azure Data Lake Analytics
- Azure SQL

## – Activities

- Copy
- HDInsight (Pig, Hive, ...)
- Azure ML Batch Scoring
- Stored Procedure
- .NET

## ■ Unterstützte Datenquellen

- Azure Blob
- Azure Table
- Azure SQL Database
- Azure SQL Data Warehouse
- Azure DocumentDB
- Azure Data Lake Store
- SQL Server on-premises/Azure IaaS
- File System On-premises/Azure IaaS
- Oracle Database On-premises/Azure IaaS
- MySQL Database On-premises/Azure IaaS
- DB2 Database On-premises/Azure IaaS
- Teradata Database On-premises/Azure IaaS
- Sybase Database On-premises/Azure IaaS
- PostgreSQL Database On-premises/Azure IaaS

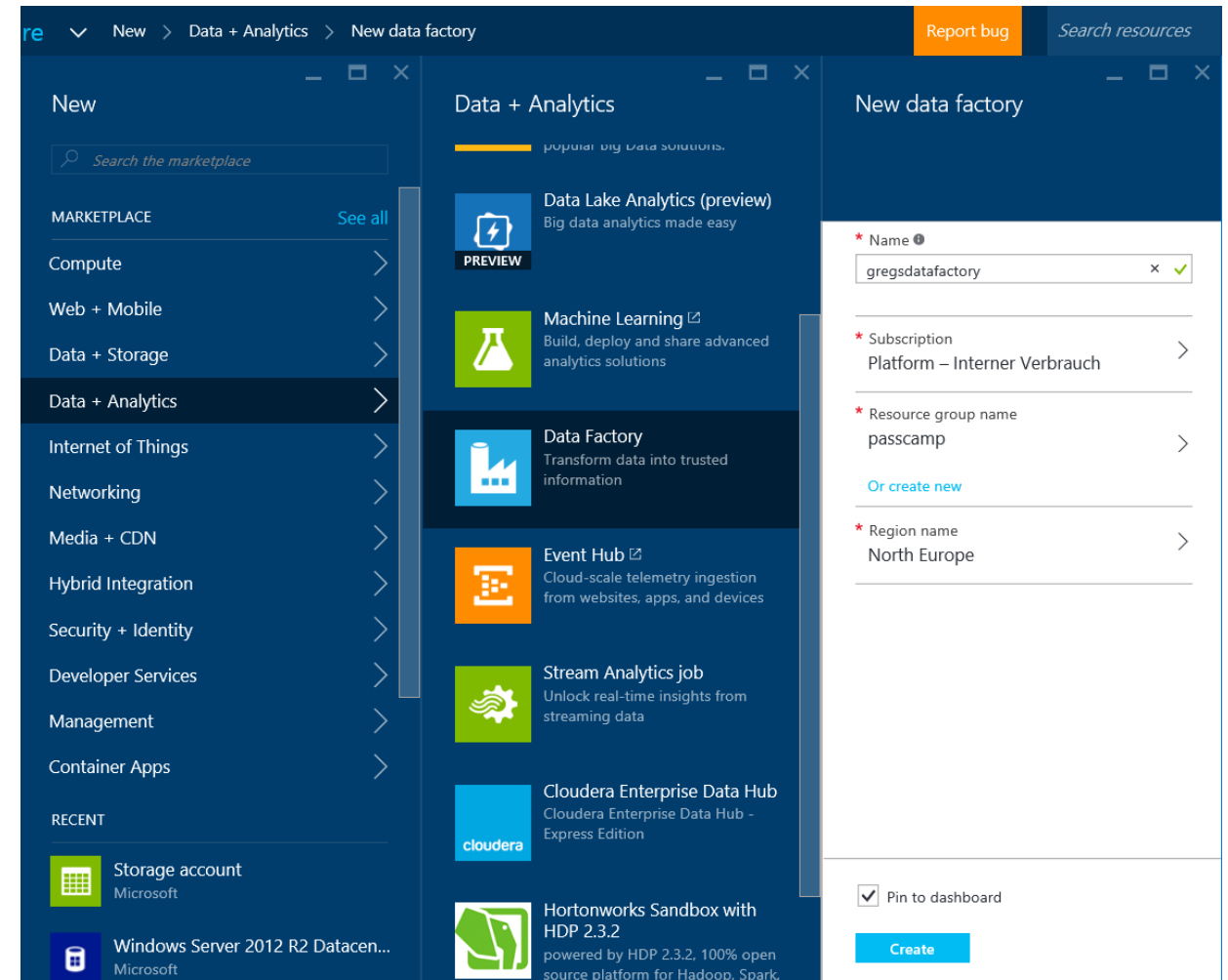
# Mit Data Factories arbeiten

# Warnung: Achtung – Work in Progress!

- „Es liegt nicht an Euch“
  - Speziell die Design-Tools haben noch Ihre Macken
    - z.B. fehlende Möglichkeit Elemente aus einem Visual Studio-Projekt zu löschen
- Verfügbarkeit
  - derzeit nur North Europe & West US
- Die Begriffe sind noch nicht „stabil“
  - z.B. Dataset (in Doku) vs. „Table“ in Visual Studio
- Die Metadaten-Schemata ändern sich hin und wieder
- „RTFM“ ist schwierig
  - die Dokumentation ist ... „übersichtlich“ und gut verteilt

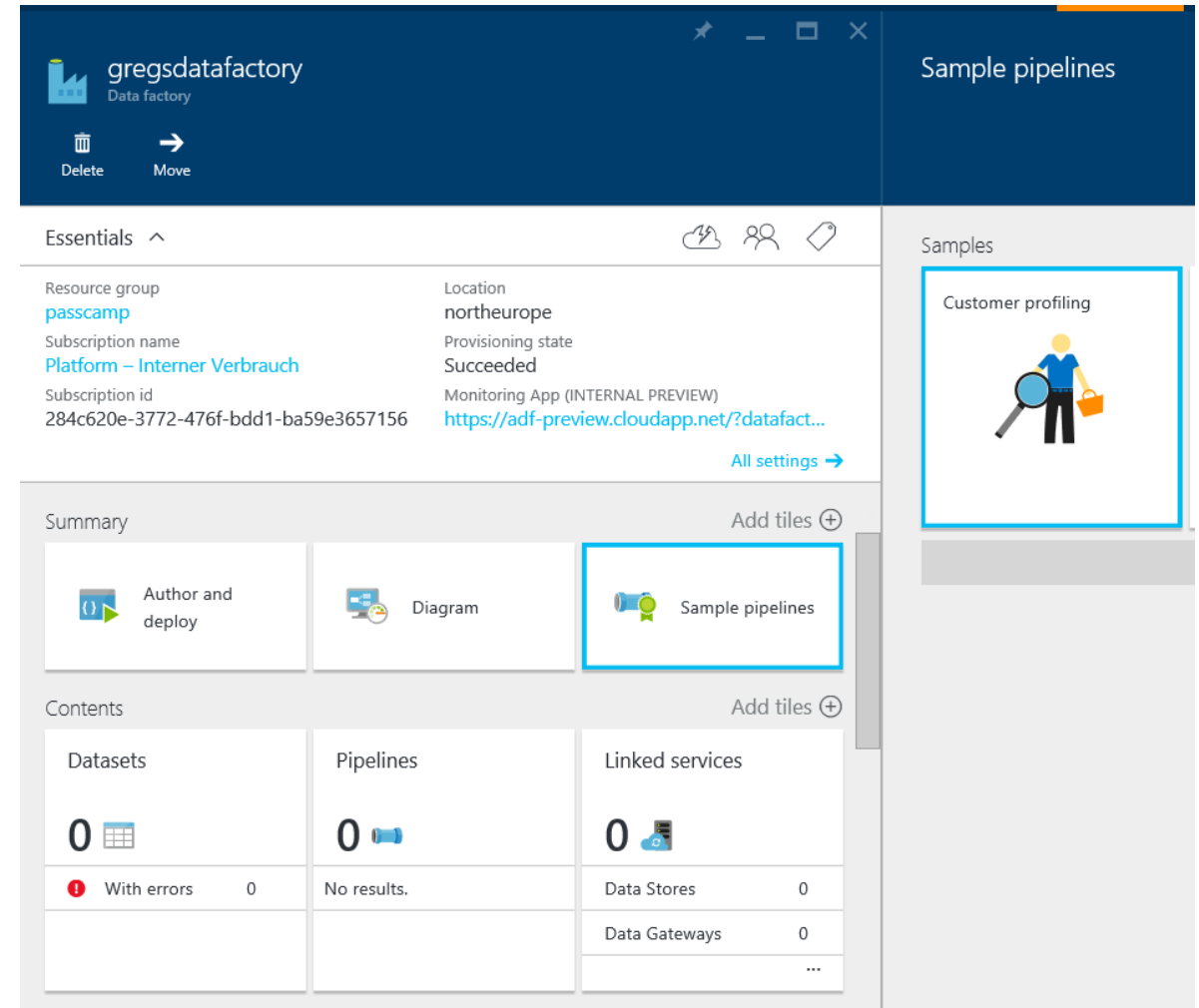
# Eine neue Data Factory anlegen

- Neues Portal
- Data > Analytics
- Ressource-Gruppen verwenden
  - für bessere Übersicht
  - einfachere Security
- Auf die Region achten!
- Die Factory ist ein Container für Pipes, Data Sets, ...



# Tipp: Erste Schritte mit den ADF Beispiel-Pipelines

- Einstieg mit fertigen Beispielen
- Ressourcen am besten selbst vorbereiten
  - gregsadfsample (Ressource Group)
  - gregsadfsamplestorage
  - ...sqlserver
  - ...sqldatabase
- Und wieder: auf Regionen achten und Ressource-Gruppen verwenden!



The screenshot shows the Azure Data Factory portal for a resource group named 'gregsdatafactory'. The interface includes a top navigation bar with 'Delete' and 'Move' actions. The main content area is divided into 'Essentials' and 'Summary' sections. The 'Essentials' section displays metadata for the resource group, including its location ('northeurope'), subscription name ('Platform – Interner Verbrauch'), and subscription ID. The 'Summary' section provides an overview of the factory's components: 0 datasets, 0 pipelines, and 0 linked services. The 'Sample pipelines' tile is highlighted with a blue border. On the right side, a 'Samples' panel shows a list of example pipelines, with 'Customer profiling' selected and highlighted.

**Essentials**

Resource group: [passcamp](#)  
Subscription name: [Platform – Interner Verbrauch](#)  
Subscription id: 284c620e-3772-476f-bdd1-ba59e3657156  
Location: northeurope  
Provisioning state: Succeeded  
Monitoring App (INTERNAL PREVIEW): <https://adf-preview.cloudapp.net/?datafact...>  
[All settings →](#)

**Summary**

**Author and deploy** **Diagram** **Sample pipelines**

**Contents**

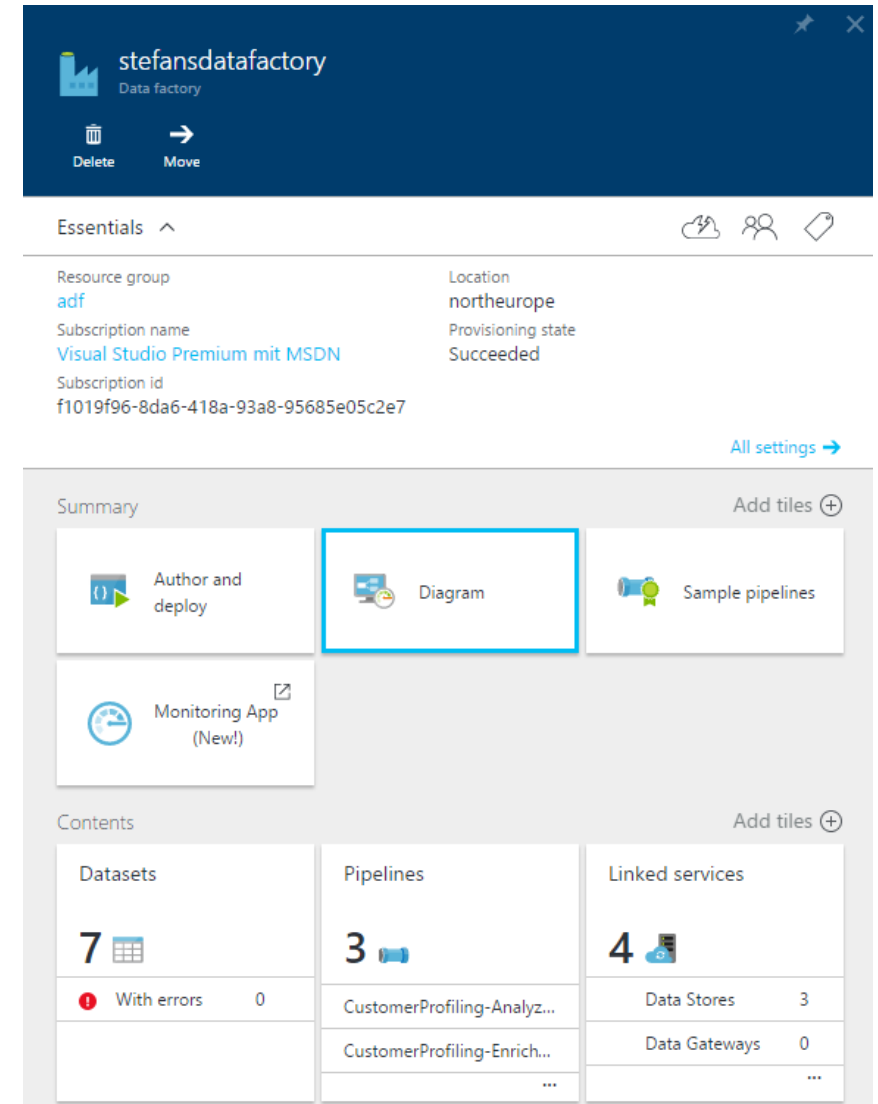
Datasets		Pipelines		Linked services	
0		0		0	
With errors	0	No results.		Data Stores	0
				Data Gateways	0
					...

**Sample pipelines**

Customer profiling

# Im Azure Portal arbeiten

- Factory-Dashboard
  - Werkzeuge für das Erstellen der Factory
  - Zustandsüberwachung
  - Manuelles Starten von Vorgängen
  - Auslastung und Diagnosen
  - **Neu! Erweitertes Monitoring**



The screenshot shows the Azure Data Factory portal interface. At the top, the header bar displays the factory name 'stefansdatafactory' and 'Data factory'. Below this, there are 'Delete' and 'Move' buttons. The main content area is divided into sections. The 'Essentials' section shows metadata: Resource group 'adf', Subscription name 'Visual Studio Premium mit MSDN', Subscription id 'f1019f96-8da6-418a-93a8-95685e05c2e7', Location 'northeurope', and Provisioning state 'Succeeded'. Below this is a 'Summary' section with three tiles: 'Author and deploy', 'Diagram' (highlighted with a blue border), and 'Sample pipelines'. A 'Monitoring App (New!)' tile is also visible. The 'Contents' section at the bottom provides a summary of resources: 7 Datasets (0 with errors), 3 Pipelines (listing 'CustomerProfiling-Analyz...' and 'CustomerProfiling-Enrich...'), and 4 Linked services (listing 'Data Stores' with 3 and 'Data Gateways' with 0).

stefansdatafactory  
Data factory

Delete Move

Essentials ^

Resource group  
adf

Subscription name  
Visual Studio Premium mit MSDN

Subscription id  
f1019f96-8da6-418a-93a8-95685e05c2e7

Location  
northeurope

Provisioning state  
Succeeded

All settings →

Summary Add tiles +

Author and deploy

Diagram

Sample pipelines

Monitoring App (New!)

Contents Add tiles +

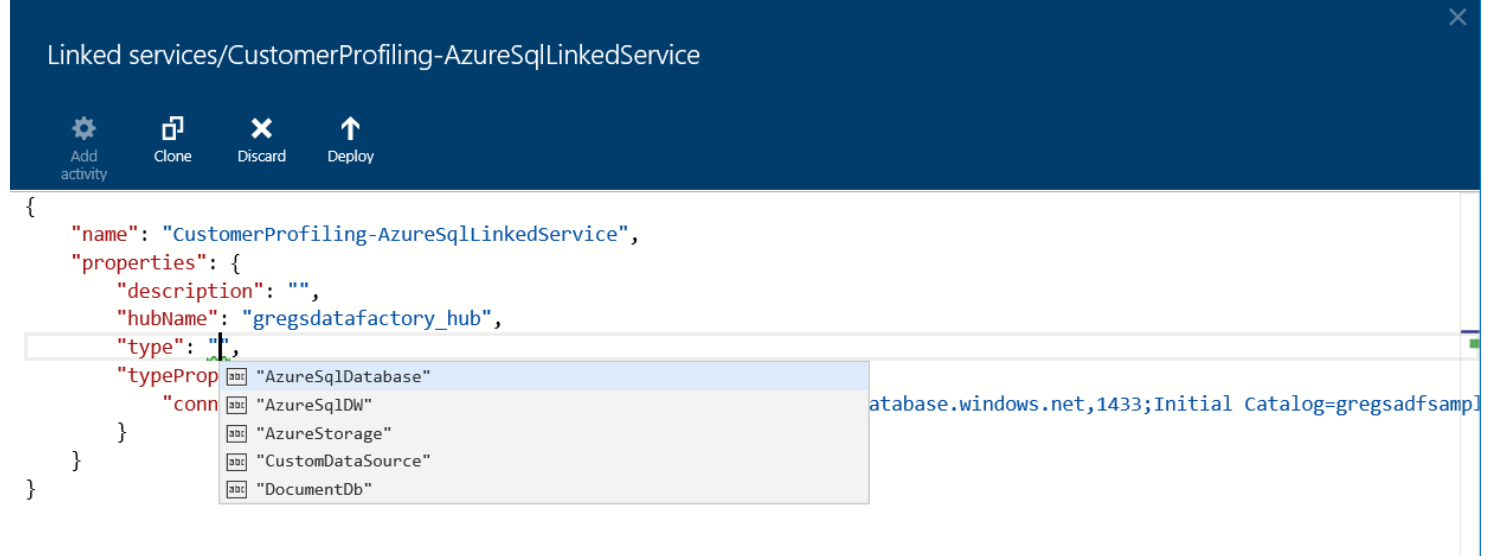
Datasets  
7  
With errors 0

Pipelines  
3  
CustomerProfiling-Analyz...  
CustomerProfiling-Enrich...  
...

Linked services  
4  
Data Stores 3  
Data Gateways 0  
...

# Ein Element entwerfen & deployen

- das Design ist textbasiert
  - JSON-Format
  - Für alle Elemente existieren Schemata
- das heißt tatsächlich häufig „Design by Copy and Paste“  
☹️



The screenshot shows the 'Linked services/CustomerProfiling-AzureSqlLinkedService' window in Azure Data Studio. The JSON configuration is as follows:

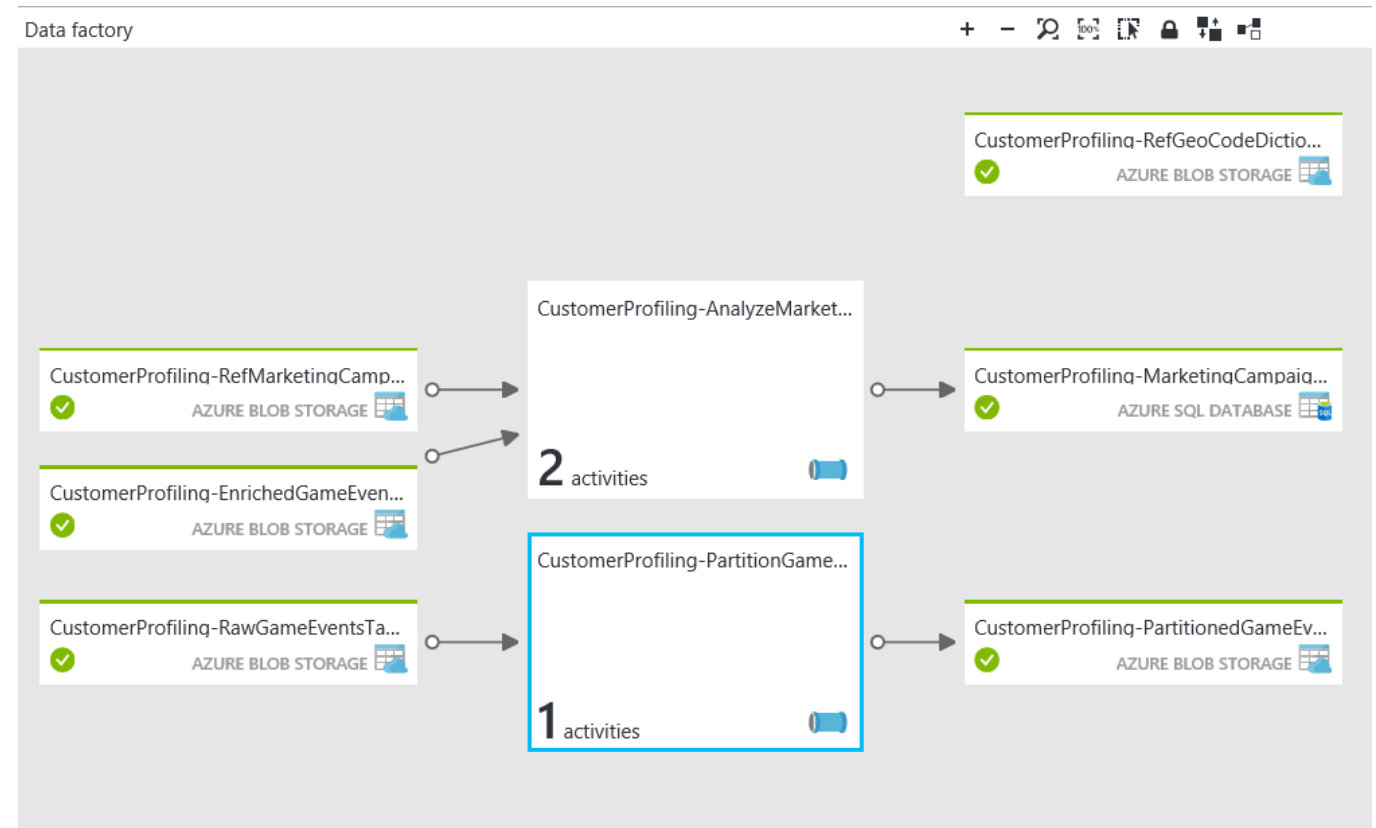
```
{
  "name": "CustomerProfiling-AzureSqlLinkedService",
  "properties": {
    "description": "",
    "hubName": "gregsdatafactory_hub",
    "type": "AzureSqlDatabase",
    "typeProperties": {
      "connectionString": "Server=tcp:gregsdatafactory.database.windows.net,1433;Initial Catalog=gregsadsample;"
    }
  }
}
```

A dropdown menu is open for the 'type' property, showing the following options:

- AzureSqlDatabase (selected)
- AzureSqlDW
- AzureStorage
- CustomDataSource
- DocumentDb

# Übersicht mit dem Diagramm gewinnen

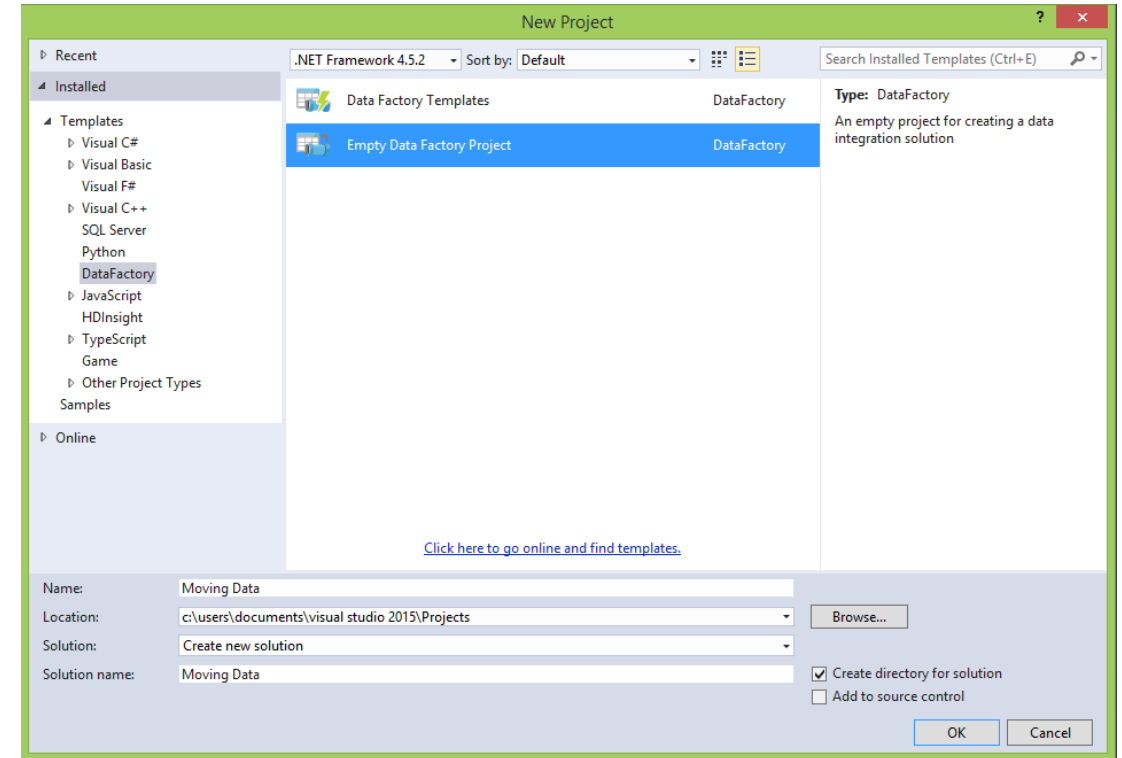
- Das Diagramm zeigt eine Übersicht über die Factory an
- Es ist *kein grafischer Designer*
- Das Lineage-Highlighting vereinfacht die Orientierung in komplexen Factories





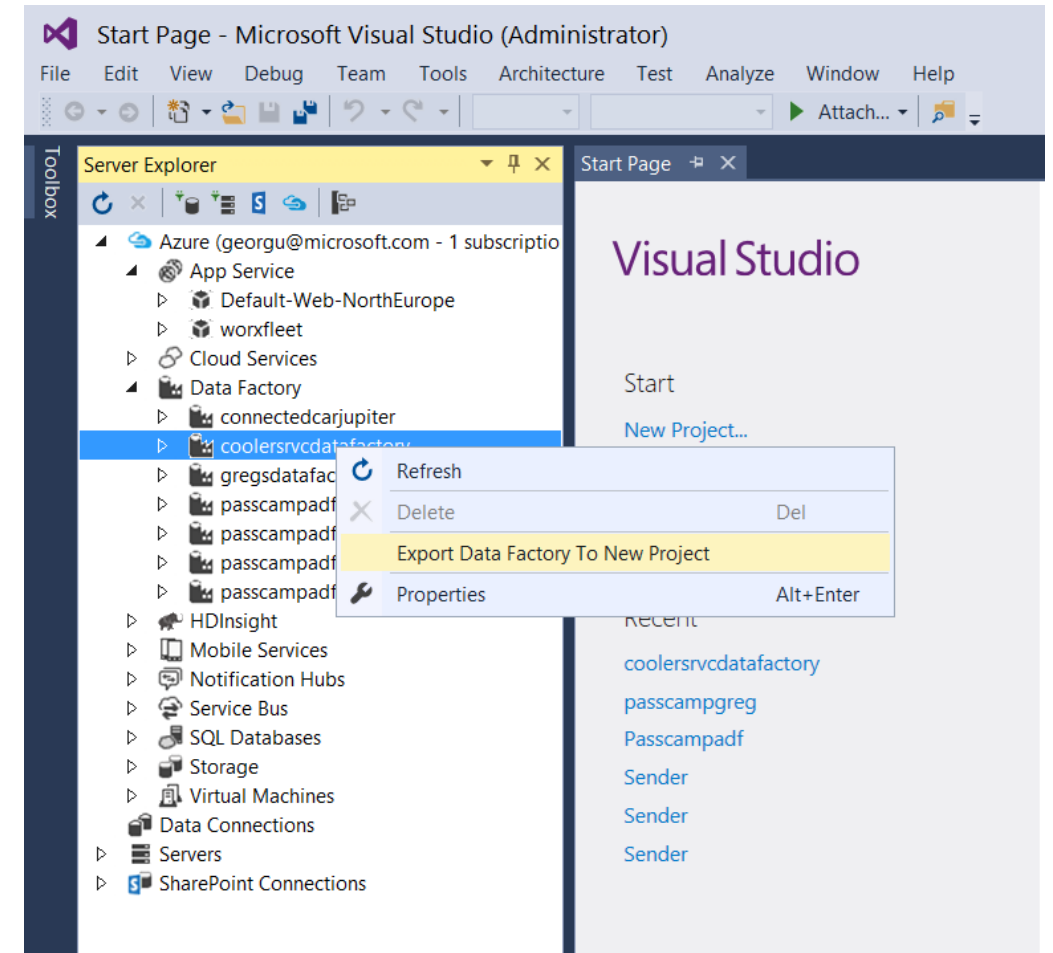
# In Visual Studio entwickeln

- Voraussetzung: Azure .NET SDK ab 2.7 aufwärts
- Visual Studio ab Version 2013
  - Syntax-Highlighting erst ab 2015
- Mit dem SDK Setup werden die passenden Projekt-Templates installiert
- Prozess:
  - Offline-Entwurf
  - Build
  - Deploy
- Monitoring findet im Portal statt
- Einbindung in Source Code-Kontrolle
- Tipp: Save & Build nach jedem neuen Element



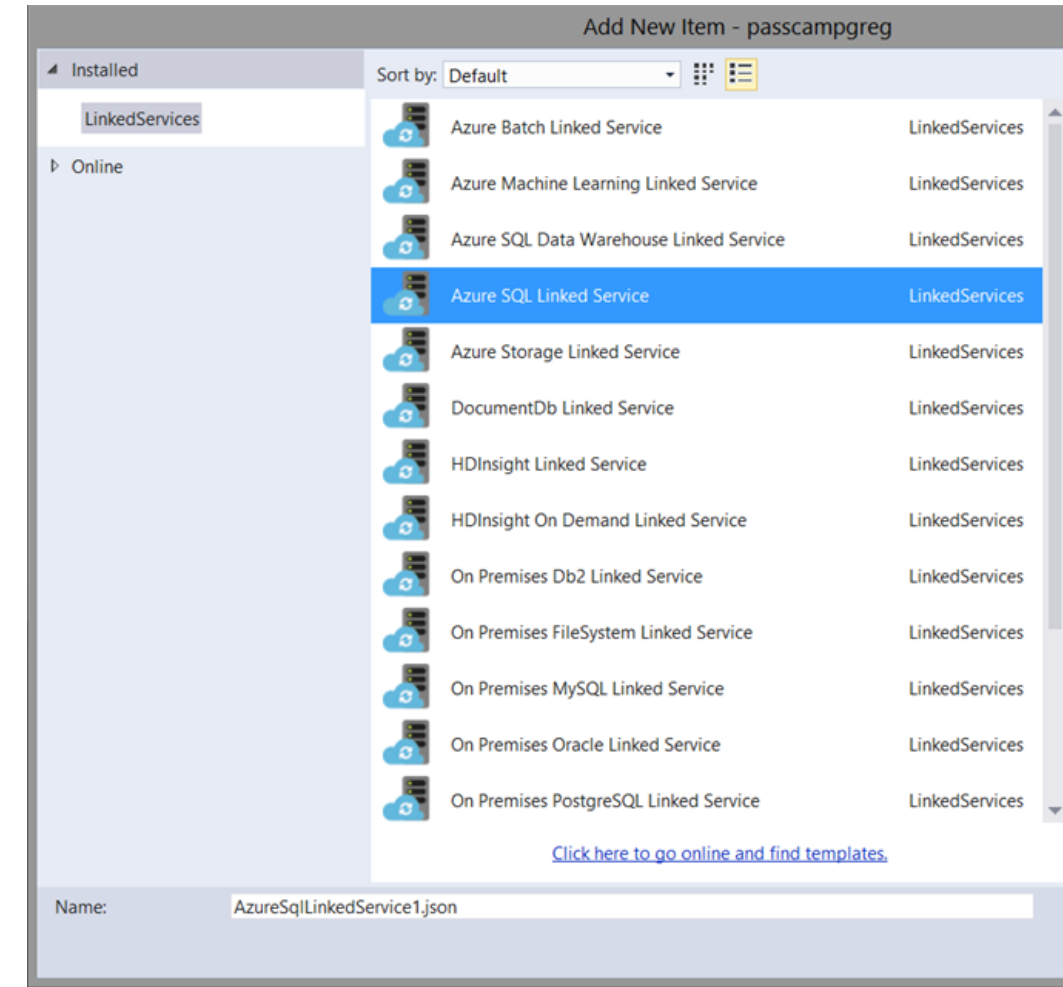
# Ein Projekt Reverse engineeren

- Im Portal vorhandene Factories können einfach in ein Projekt überführt werden
- Server Explorer
  - Export Data Factory to New Project



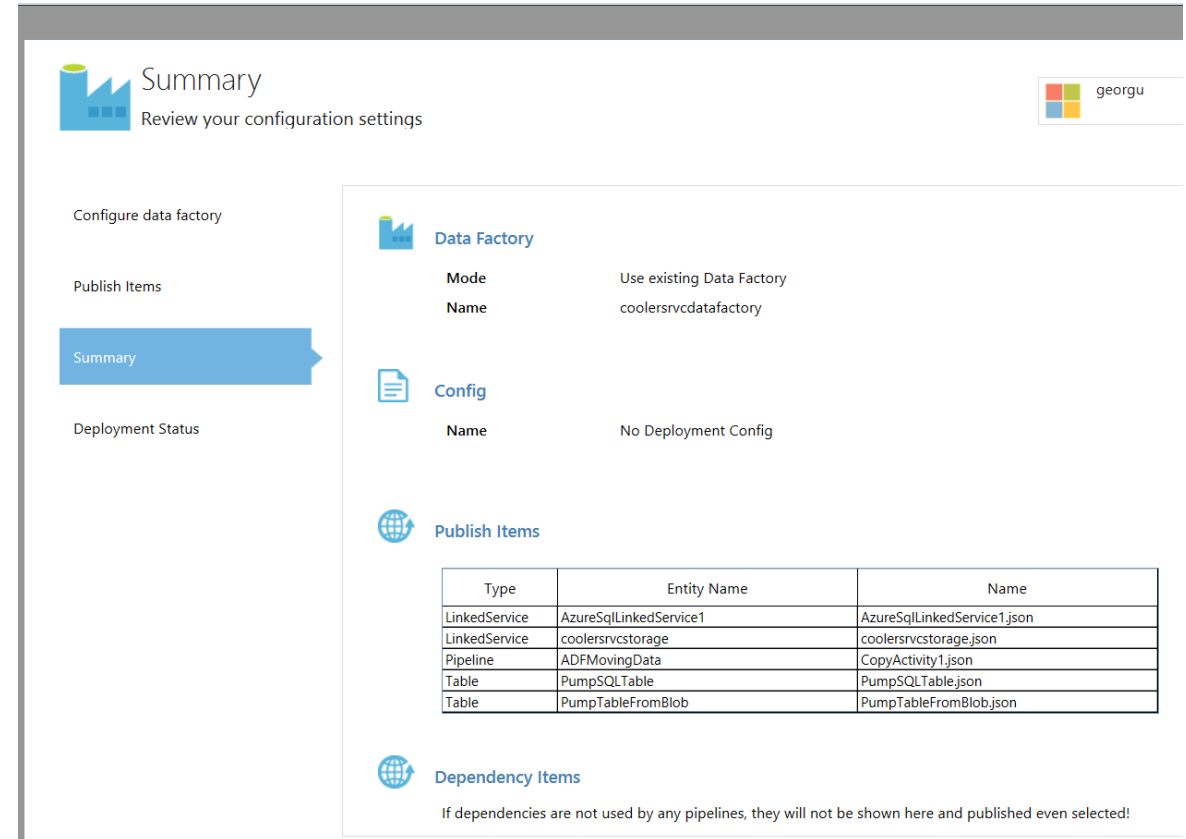
# Neue Objekte entwerfen

- Templates enthalten
  - Code-Fragmente
  - Verweis auf das passende Schema



# Von Visual Studio aus publizieren

- in neue oder bestehende Factory publizieren
- Offline-Projekt wird mit der Online-Factory verglichen
  - zu publizierende oder zu löschende Elemente können an- oder abgewählt werden



The screenshot shows the 'Summary' page for an Azure Data Factory project. The page title is 'Summary' with the subtitle 'Review your configuration settings'. The user 'georgu' is logged in. The left sidebar contains navigation links: 'Configure data factory', 'Publish Items', 'Summary' (highlighted), and 'Deployment Status'. The main content area displays the following information:

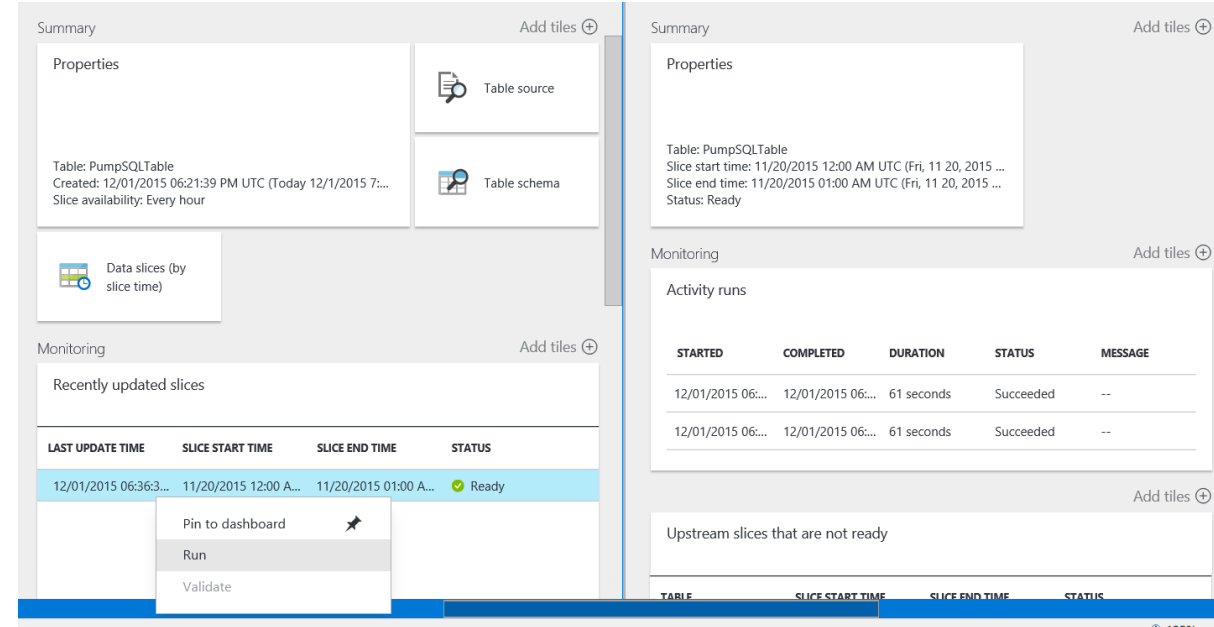
- Data Factory**: Mode is 'Use existing Data Factory', Name is 'coolersrvdatafactory'.
- Config**: Name is 'No Deployment Config'.
- Publish Items**: A table listing items to be published.
- Dependency Items**: A section with a note about dependencies.

Type	Entity Name	Name
LinkedService	AzureSqlLinkedService1	AzureSqlLinkedService1.json
LinkedService	coolersrvstorage	coolersrvstorage.json
Pipeline	ADFMovingData	CopyActivity1.json
Table	PumpSQLTable	PumpSQLTable.json
Table	PumpTableFromBlob	PumpTableFromBlob.json

If dependencies are not used by any pipelines, they will not be shown here and published even selected!

# Ausführung überwachen

- im Portal ADF Dashboard unter *Contents*
- am besten unter Datasets nachschauen
- Infos:
  - aktueller Status
  - ausgeführte Aktivitäten
  - Meldungen
  - ..
- in den *Slices* kann Aktivität neu gestartet werden
- Jetzt gibt's aber auch die „Monitoring App“



The screenshot displays the Monitoring App interface for a dataset named 'PumpSQLTable'. It is divided into two main panels: 'Summary' and 'Monitoring'.

**Summary Panel:**

- Properties:** Table: PumpSQLTable, Created: 12/01/2015 06:21:39 PM UTC (Today 12/1/2015 7:..., Slice availability: Every hour.
- Table source** and **Table schema** links are available.
- Data slices (by slice time)** section.

**Monitoring Panel:**

- Recently updated slices** table:

LAST UPDATE TIME	SLICE START TIME	SLICE END TIME	STATUS
12/01/2015 06:36:3...	11/20/2015 12:00 A...	11/20/2015 01:00 A...	Ready

A context menu is open over the 'Ready' status, showing options: 'Pin to dashboard', 'Run', and 'Validate'.

- Activity runs** table:

STARTED	COMPLETED	DURATION	STATUS	MESSAGE
12/01/2015 06:...	12/01/2015 06:...	61 seconds	Succeeded	--
12/01/2015 06:...	12/01/2015 06:...	61 seconds	Succeeded	--

**Upstream slices that are not ready** section.

# Überwachen und verwalten mit der Monitoring App

- Besteht aus 3 Teilen:
  - Resource Explorer
  - Monitoring Views
  - Alerts
- Debugging der Data Factory über versch. Ansichten
- Starten von Batches mit abh. Slices
- Anlegen von Email Alerts (detaillierter als bisher)

Summary				Monitor	
Linked services	3	Datasets	4	Metrics	
DATA STORES	3	Pipelines	4	SUCCEEDED (02/15/2016-02/22/2016)	28 ✓
GATEWAYS	0			FAILED (02/15/2016-02/22/2016)	0 ✗
				IN PROGRESS	0 ⚙
				View recent activity windows	
				View failed activity windows	
				View in-progress activity windows	

Data factory

RESOURCE EXPLORER

- Data Factories
  - stefansdatafactory
    - Pipelines
      - CustomerProfiling-AnalyzeMarketingCampaignPipeline
      - CustomerProfiling-EnrichGameLogsPipeline
      - CustomerProfiling-PartitionGameLogsPipeline
    - Datasets
      - CustomerProfiling-EnrichedGameEventsTable
      - CustomerProfiling-MarketingCampaignEffectivenessBlobTable
      - CustomerProfiling-MarketingCampaignEffectivenessSQLTable
      - CustomerProfiling-PartitionedGameEventsTable
      - CustomerProfiling-RawGameEventsTable
      - CustomerProfiling-RefGeoCodeDictionaryTable
      - CustomerProfiling-RefMarketingCampaignTable
    - Linked Services
      - CustomerProfiling-AzureSqlLinkedService
      - CustomerProfiling-HDInsightLinkedService
      - CustomerProfiling-HDInsightStorageLinkedService
      - CustomerProfiling-StorageLinkedService
    - Gateways

Data factory

MONITORING VIEWS

- System Views
  - Recent activity windows
  - Failed activity windows
  - In-progress activity windows

Data factory

ALERTS

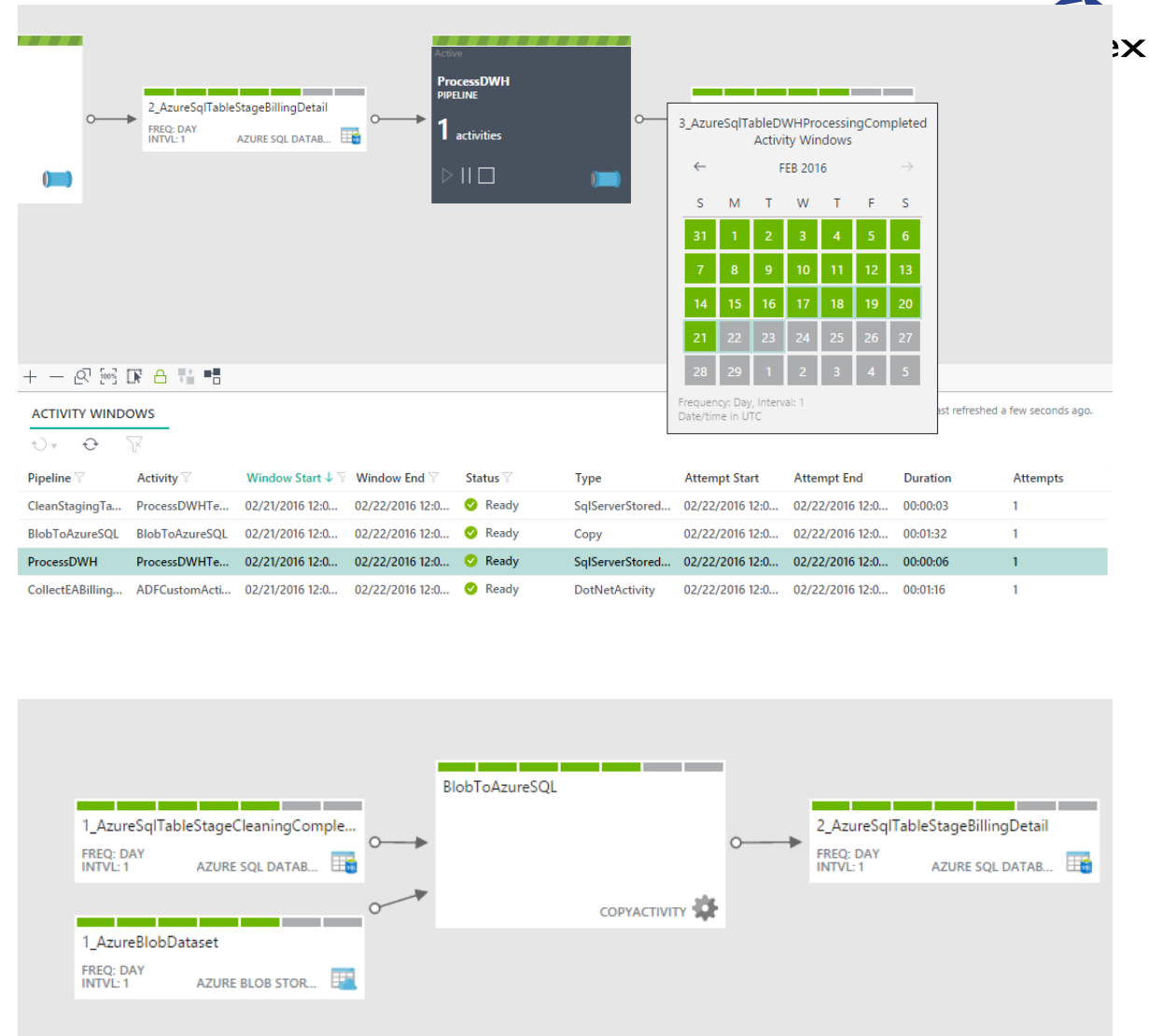
+ Add Alert

Obacht



# Resource Explorer

- Detaillierte Auswertungen über Verläufe, schön dargestellt
- Drill down in die Pipelines



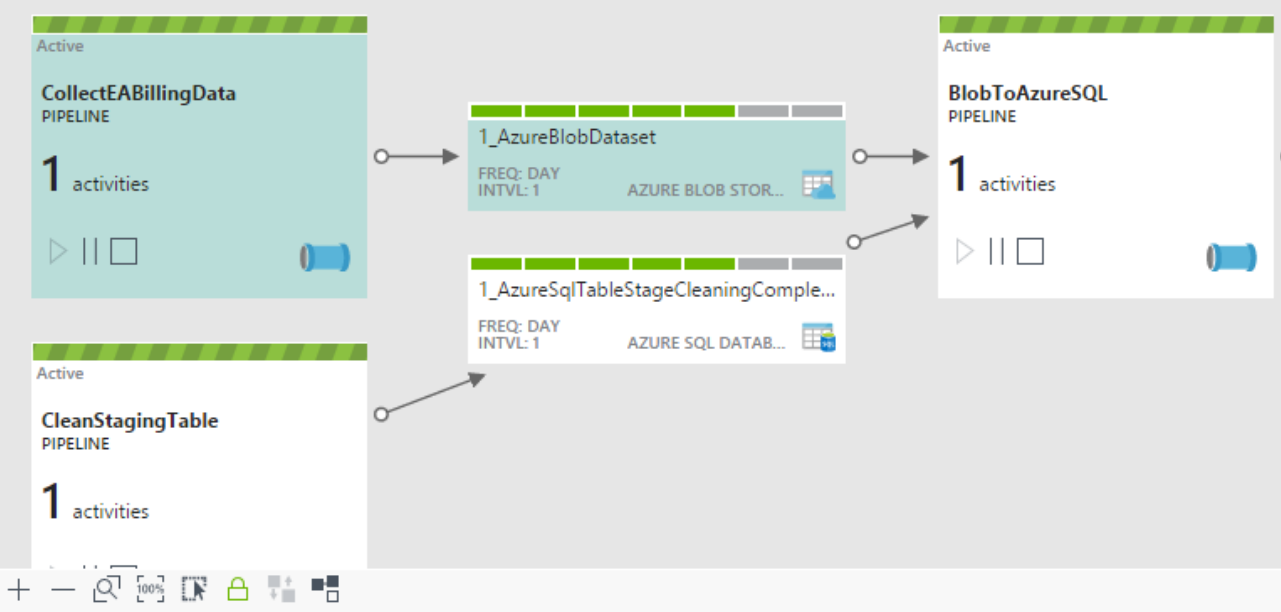


# Resource Explorer zum Restart

**Data Factories**

- azurebillingdatafactory
  - Pipelines
    - BlobToAzureSQL
    - CleanStagingTable
    - CollectEABillingData
    - ProcessDWH
  - Datasets
    - 1\_AzureBlobDataset
    - 1\_AzureSqlTableStageCl...
    - 2\_AzureSqlTableStageBi...
    - 3\_AzureSqlTableDWHPr...
  - Linked Services
    - AzureBatchLinkedService
    - AzureSqlLinkedService
    - StorageLinkedService
  - Gateways

Start time (UTC): 11/22/2015 04:10 pm End time (UTC): 02/23/2016 04:10 pm Apply



```

graph LR
    A[CollectEABillingData PIPELINE] --> B[1_AzureBlobDataset]
    B --> C[1_AzureSqlTableStageCleaningComple...]
    C --> D[BlobToAzureSQL PIPELINE]
    
```

**ACTIVITY WINDOWS**

Activity	Window Start	Window End	Status	Type	Attempt Start
ProcessDWHTe...	11/25/2015 12:00...	11/26/2015 12:00...	Failed	SqlServerStored...	01/15/2016 5:33 ...
ProcessDWHTe...	11/24/2015 12:00...	11/25/2015 12:00...	Failed	SqlServerStored...	01/15/2016 5:32 ...

Rerun

Rerun with null upstream in pipeline

# Detaillierte Alerts setzen

CREATE ALERT / DETAILS
 ✕

Name

Description (optional)

Data factory
 

1 Details
 2 Filters
 3 Recipients

CREATE ALERT / FILTERS
 ✕

Event

Status

Substatus (optional)
 

--
 --
 Failed Resource Allocation
 Failed Execution
 Timed Out
 Failed Validation
 Abandoned

Data factory
 

1 Details
 2 Filters
 3 Recipients

CREATE ALERT / RECIPIENTS
 ✕

☒ Email subscription admins

Additional administrator email

# Kosten

# Was kostet denn das eigentlich?

...gar nicht so einfach zu beantworten

- „Die ersten 5 Aktivitäten mit niedriger Auslastung sind ... in einem Monat kostenlos“ - ah, ja...
- Häufigkeitsklassen
  - max. 1 X pro Tag: „weniger häufig“. Sonst: „sehr häufig“ (also z.B. 2 X pro Tag ;-)
- Kosten für Aktivitäten

	Weniger häufig	Sehr häufig
Cloud	€0,506	€0,6747
On-Premises	€1,26	€2,1083

- ab 100 Aktivitäten/Monat 20% Rabatt
- Data Movement Service (wird in der Quelle gemessen)
  - Cloud: €0.2109/h
  - On-Premises: €0.0844/h

# Beispiel-Rechnung (Customer Profiling-Szenario)

- Aktivitäten / Monat
  - 30 X Copy – 4: 28 On-Prem
  - 4 X Join/Aggregate – 4: 0 Cloud
- Data Movement
  - User aus dem on Prem SQL Server holen: 30 X 1h
  - User aus dem Blob Storage holen: 4 X 0,25h
  - Game-Daten aus dem Blob Storage holen: 4 X 5h

Aktivitäten On-Prem	28	€1,26	€35,00
Movement On-Prem	30	€0.0844	€2,50
Movement Cloud	21	€0.2109	€4,40
			ca. €42/Monat

- Pro Jahr ca. 500€ für die Factory

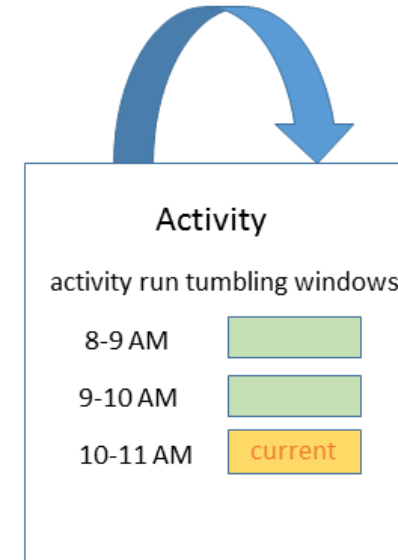
# Vertiefungen

- start/end-Elemente in der Pipeline
  - Pipeline steht in diesem Zeitraum prinzipiell zur Verfügung
- Slice
  - Zeitscheibe (Intervall) für die Dataset-Verarbeitung
- Frequenz der Generierung kann im Input-Dataset, einer Aktivität, im Output-Dataset definiert sein
  - das Output-Dataset gibt den Takt vor
  - Aktivität startet, wenn passendes Input Slice vorhanden ist
- Bezeichnung der Zeitplanungs-Elemente
  - Dataset: Availability
  - Aktivität: Scheduler
  - Aufbau dieser Element-Typen ist identisch

# Zeitplanungs-Eigenschaften

- "scheduler":
  - { "frequency": "Hour", "interval" }
  - Intervalle starten um 0 Uhr
- style: StartOfInterval / EndOfInterval
  - frequency = Month & style = EndOfInterval:  
Letzter Tag im Monat
- anchorDateTime: Zeitpunkt, von dem  
aus slices berechnet werden
  - z.B. 2015-10-19T08:00:00" }

Scheduler: Run every hour

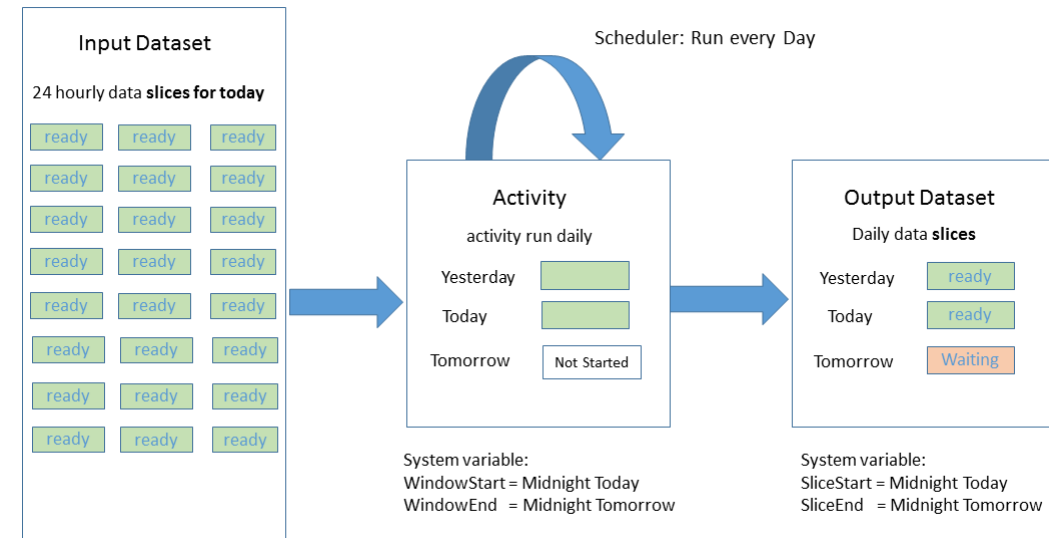
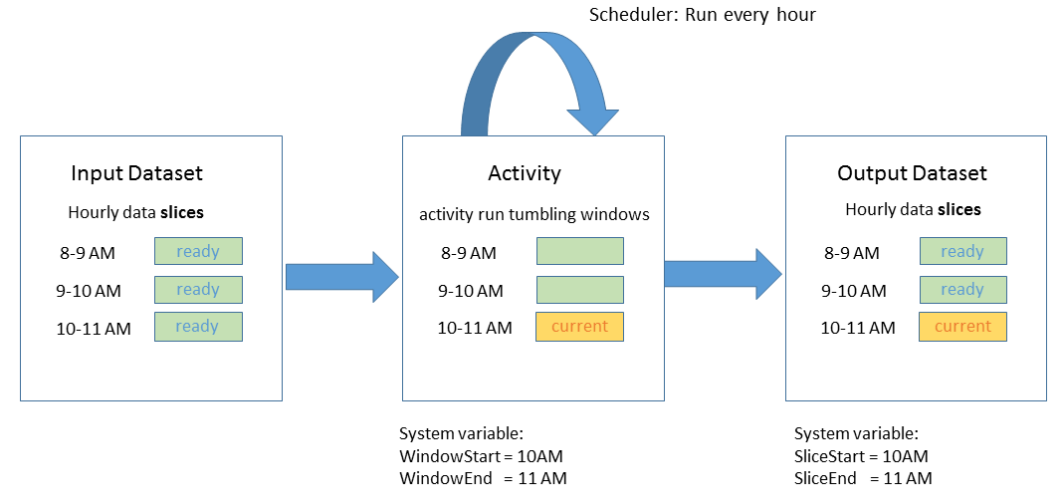


System variable:  
WindowStart = 10AM  
WindowEnd = 11 AM



# Zeitplanung in einer Pipeline

- slices können unterschiedlich sein
- Output-Slices und Aktivitäts-Windows müssen übereinstimmen
- die aktuellen Zeitpunkte können über Systemvariablen abgefragt werden
- Startdatum in der Vergangenheit
  - alle Slices werden automatisch aufgefüllt



# Beispiel: SQL Dataset als Input & Blob als Output

- "availability": { "frequency": "Hour",  
"interval": 1 },
- soll zwischen 8h und 11h ergeben:
  - mypath/2015/1/1/8/Data.txt  
10002345,334,2,2015-01-01 08:24:00.3130000  
10002345,347,15,2015-01-01 08:24:00.6570000  
10991568,2,7,2015-01-01 08:56:34.5300000
  - mypath/2015/1/1/9/Data.txt  
10002345,334,1,2015-01-01 09:13:00.3900000  
24379245,569,23,2015-01-01 09:25:00.3130000  
16777799,21,115,2015-01-01 09:47:34.3130000

CustomerID	ProductID	Units	timestampcolumn
10002345	334	2	2015-01-01 08:24:00.313
10002345	347	15	2015-01-01 08:24:00.657
10991568	2	7	2015-01-01 08:56:34.530
10002345	334	1	2015-01-01 09:13:00.390
24379245	569	23	2015-01-01 09:25:00.313
16777799	21	115	2015-01-01 09:47:34.313

# Beispiel: Copy Activity für time sliced-SQL Daten

```
...
"activities": [{
  "type": "Copy", "name": "AzureSQLtoBlob", "description": "copy activity",
  "typeProperties": {
    "source": { "type": "SqlSource",
    "sqlReaderQuery": "$$Text.Format(
    'SELECT * from MyTable
    WHERE timestampcolumn >= \\{0:yyyy-MM-dd HH:mm}\\'
    AND timestampcolumn < \\{1:yyyy-MM-dd HH:mm}\\' ',
    WindowStart, WindowEnd )" },
    "sink": { "type": "BlobSink", "writeBatchSize": 100000, "writeBatchTimeout": "00:05:00" } },
  "inputs": [ { "name": "AzureSQLInput" } ],
  ...
}
```

**Fazit: Mächtig aber auch strange...hier soll ein Update kommen!**

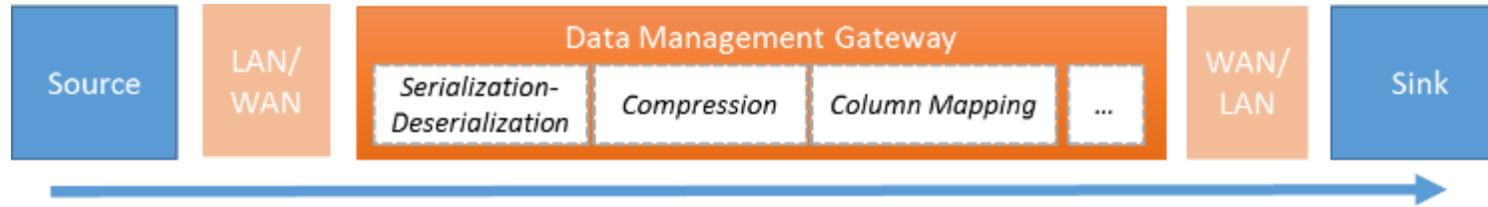
# Pig & Hive einsetzen

- HDInsight als Compute Environment
  - On-demand Cluster
  - BYOC: „Bring your own Cluster“ – bestehender Cluster
- On Demand Cluster als Linked Service

```
{ "name": "HDInsightOnDemandLinkedService",  
  "properties": { "type": "HDInsightOnDemand",  
    "typeProperties": { "clusterSize": 4,  
      "timeToLive": "00:05:00", "version": "3.1", "osType": "linux",  
      "linkedServiceName": "MyBlobStore" ...
```

# Hybride Szenarien

## – Data Management Gateway

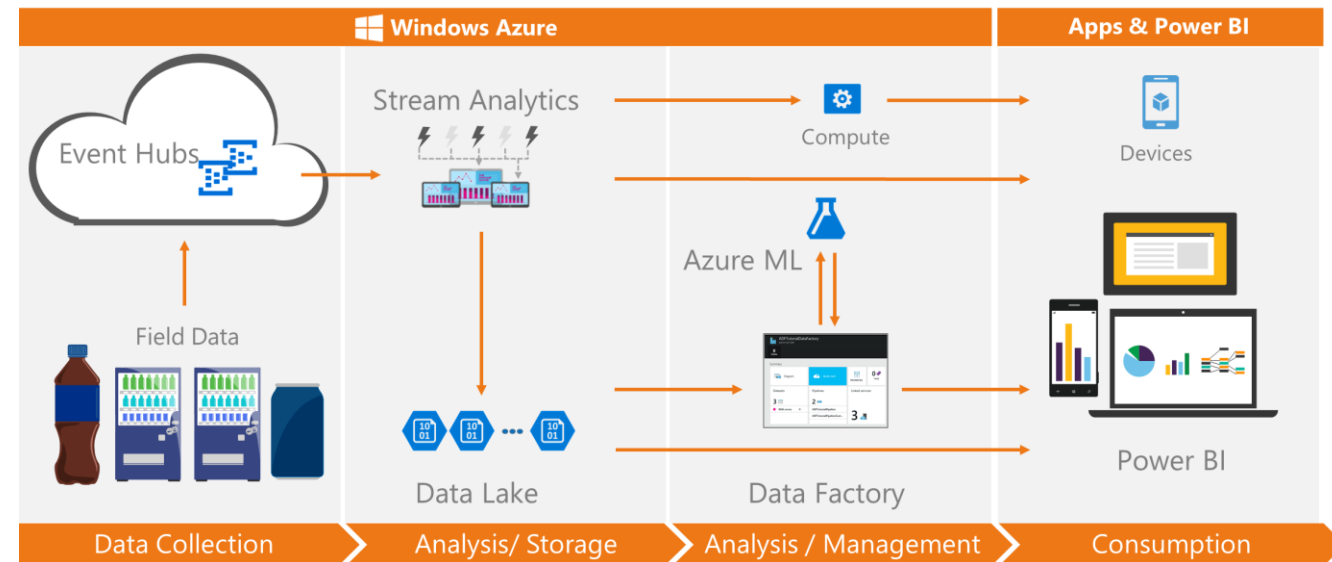


## – Erhältlich im Microsoft Download Center

- <http://www.microsoft.com/de-de/download/details.aspx?id=39717>
- läuft nur unter Windows
- max. ein Gateway auf einem Rechner
- wird auch von Power BI verwendet
- Datenquellen können sich ein Gateway teilen
- ...aber ein Gateway gehört zu einer Factory

# Vorhersagepipelines mit Azure ML

- Verwendet veröffentlichtes Azure ML Scoring Modell
- Scoring via Batch-API
- Auch Re-Training ist möglich



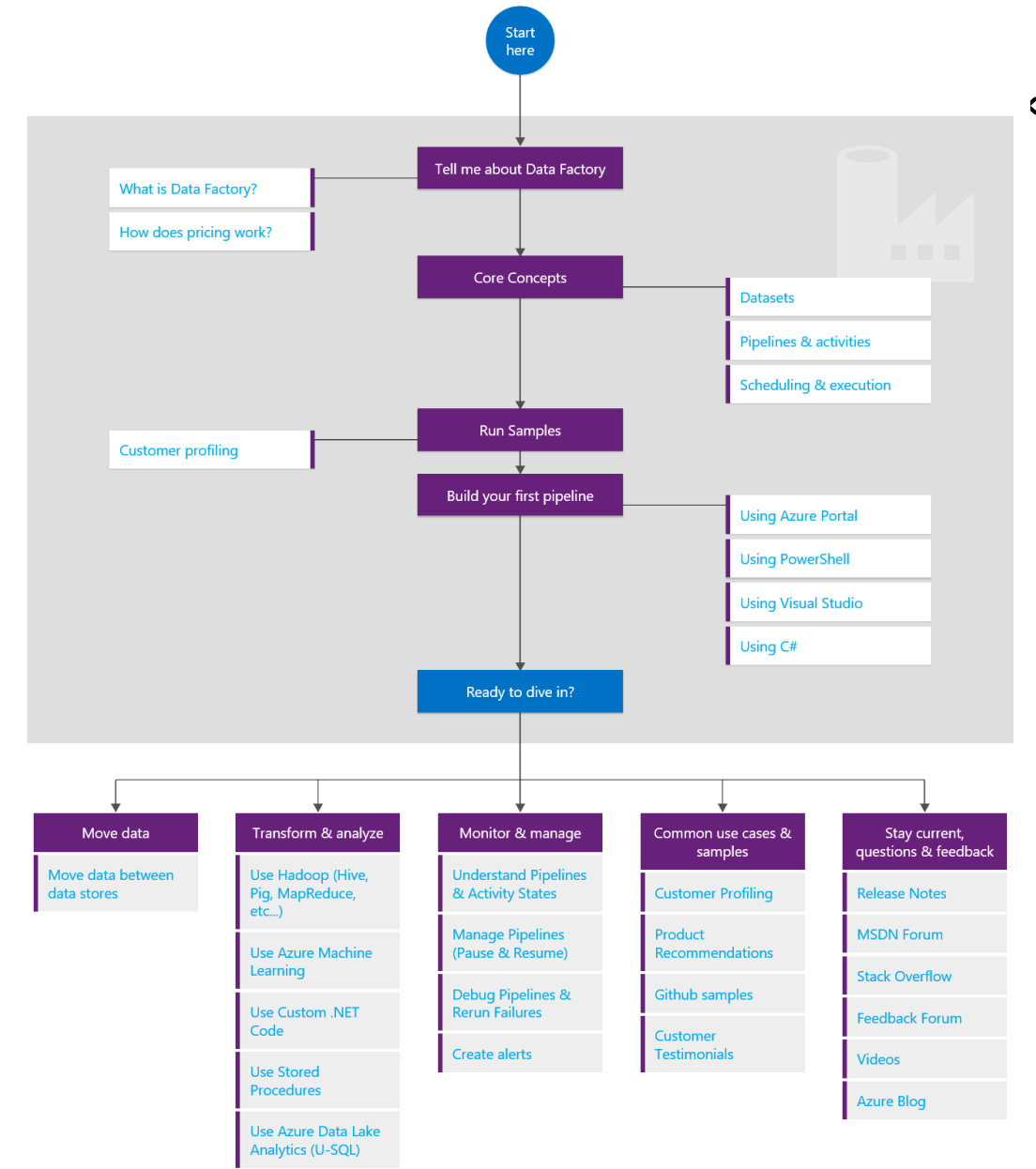
# Outlook & Call to action

# Roadmap Stand Dez. 2015

	Added Recently	Roadmap
Authoring	<ul style="list-style-type: none"><li>Web-based editor</li></ul>	<ul style="list-style-type: none"><li>Visual Studio (intellisense &amp; pipeline templates) 😊</li><li>Application templates (customer churn, recommendations, etc) 😊</li><li>Visual authoring</li></ul>
Data Movement	<ul style="list-style-type: none"><li>On premises: Oracle DB, file shares</li></ul>	<ul style="list-style-type: none"><li>Data sources added each month</li><li>Azure Document Db, Azure Search, Azure SQL DW, Azure Data Lake 😊</li></ul>
Data Production	<ul style="list-style-type: none"><li>Azure Machine Learning</li><li>HDInsight enhancements</li><li>Azure Batch</li></ul>	<ul style="list-style-type: none"><li>Additional Activities</li><li>Reference Data enhancements</li></ul>
Data Management	<ul style="list-style-type: none"><li>Monitoring diagram: lineage views, custom layout</li><li>"Recently updated datasets" view</li></ul>	<ul style="list-style-type: none"><li>Monitoring enhancements: new views, customization, resource utilization views, etc. 😊</li><li>Enhanced alerting 😊</li></ul>
Extensibility	<ul style="list-style-type: none"><li>Extension SDK -&gt; limited preview</li></ul>	<ul style="list-style-type: none"><li>Extension SDK release</li></ul>
Geo Location	<ul style="list-style-type: none"><li>Note: can orchestrate/schedule/monitor resources in all geo-regions now</li></ul>	<ul style="list-style-type: none"><li>Additional geo-regions</li></ul>



# Learning Path



<https://azure.microsoft.com/en-us/documentation/learning-paths/data-factory/>

# Call to Action

- Documentation: [azure.com/df](https://azure.com/df)
- [Samples on GitHub](#)
- Ask questions: [MSDN Forum](#)
- [Request & vote on new features](#)
- Financial services [blueprint](#)
- Case Studies
  - [Milliman - Actuarial Automation](#)
  - [Rockwell Automation - Operational Excellence](#)
  - [Ziosk – Improved Guest Experience & Satisfaction](#)



- ADF Blog  
<https://azure.microsoft.com/en-us/blog/tag/azure-data-factory/>
- ADF vs. SSIS (PASS Präsentation)  
[www.sqlpass.org/EventDownload.aspx?suid=3583](http://www.sqlpass.org/EventDownload.aspx?suid=3583)
- Channel 9 „Data Exposed“ Videos  
<https://channel9.msdn.com/Shows/Data-Exposed>

## Kontakt

Stefan Kirner  
*Head of BI Solutions*

inovex GmbH  
Ludwig-Erhard-Allee 6  
76133 Karlsruhe

Mobil: 0173 3181012  
Mail: [stefan.kirner@inovex.de](mailto:stefan.kirner@inovex.de)

