

Alexander Klein

ETL meets Azure



Who am I?

Independent BI Consultant

> 15 years experience of SQL Server

Focus on Microsoft BI Stack & AI & Azure

✉ a.klein@consulting-bi.de

🐦 @SQL_Alex

🏠 consulting-bi.de



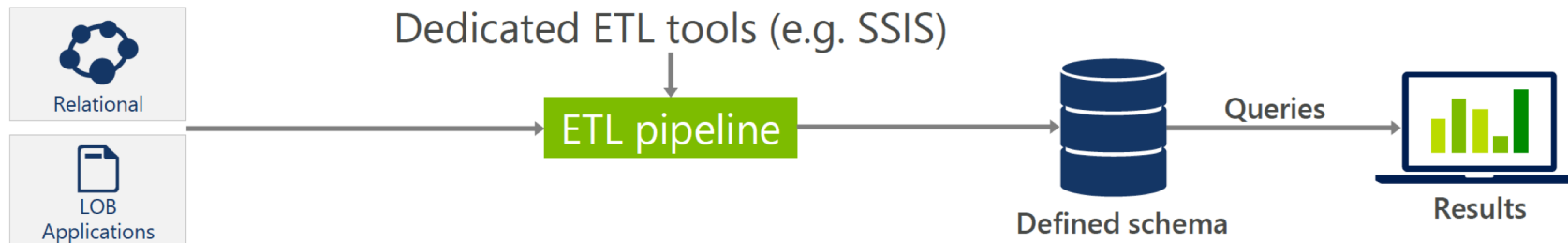
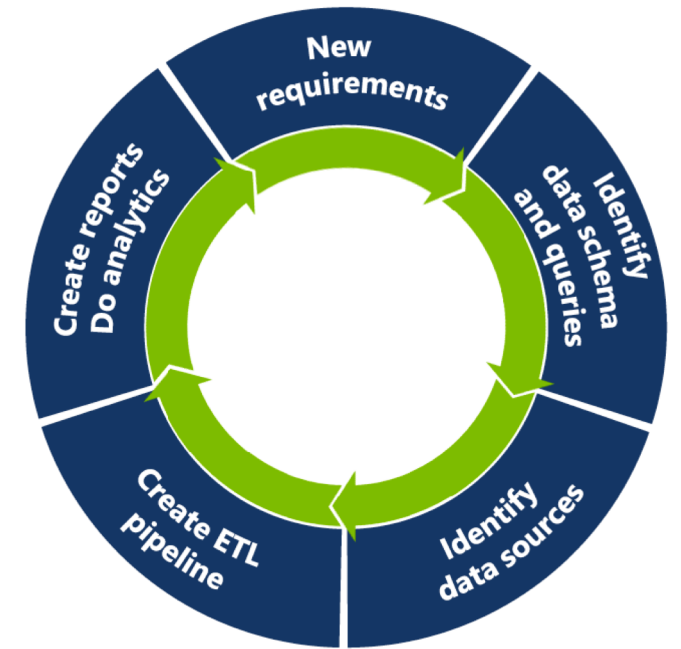
Next 60 Minutes

- ETL
- Azure Logic App
- Azure Function
- Azure Data Factory
- Azure Data Lake
- Azure Stream Analytics
- Azure Automation / Runbook



Traditional business analytics process

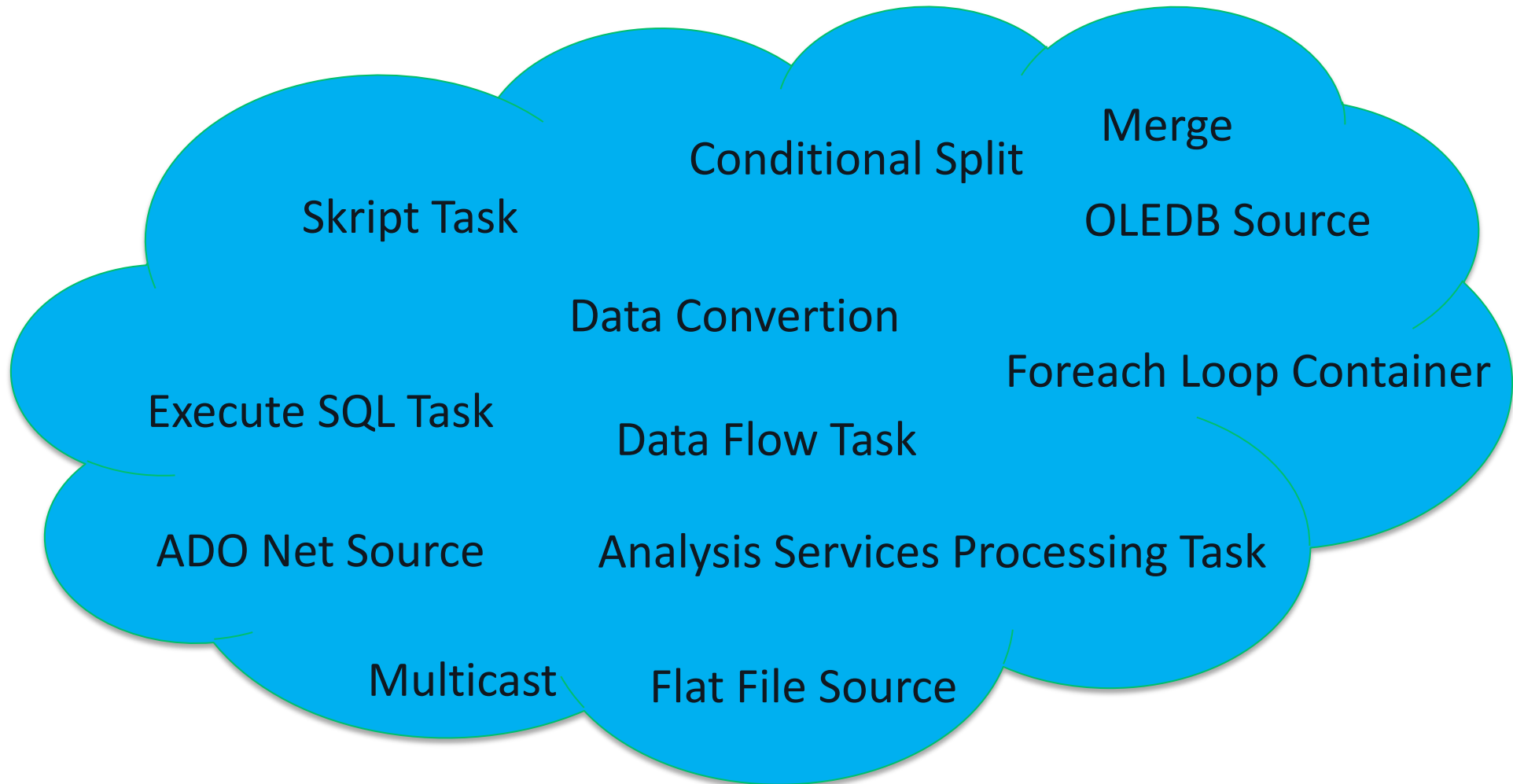
1. Start with end-user requirements to identify desired reports and analysis
2. Define corresponding databases schema and queries
3. Identify the required data source
4. Create a Extract-Transform-Load (ETL) pipeline to extract required data (curation) and transform it to target schema
5. Create reports and analyze data



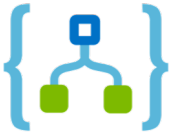
All data not immediately required is discarded or archived



On Prime (classic)



Azure Logic App



iPaaS (integration Platform as a Service)

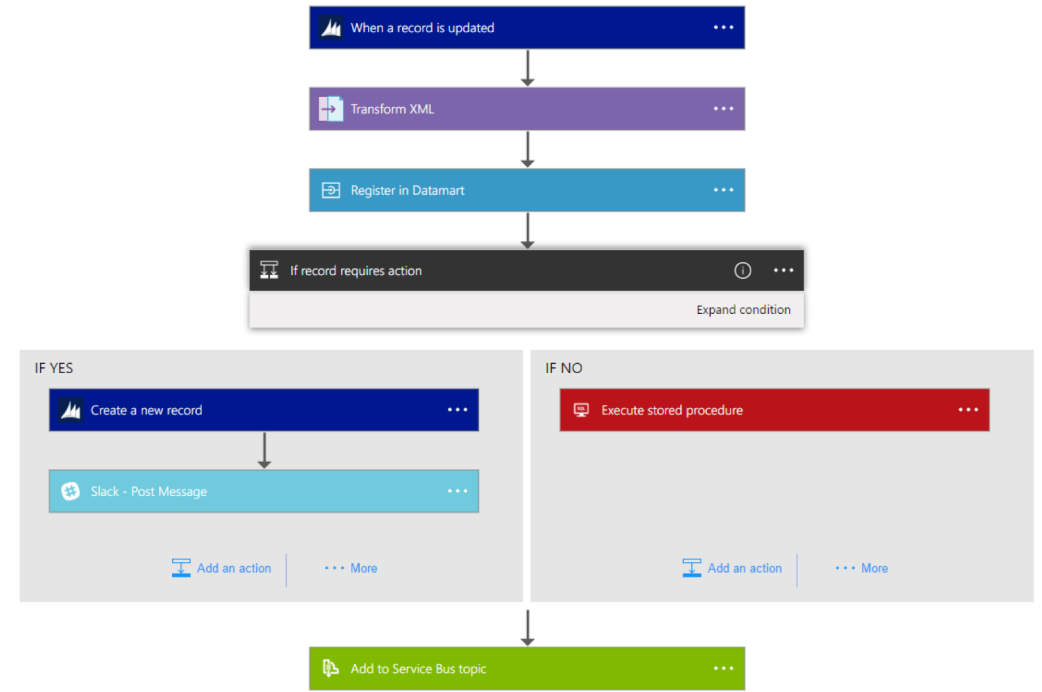
Workflow

Connectors

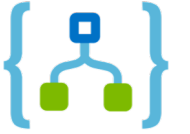
Trigger

Actions

Enterprise Integration Pack



Azure Logic App

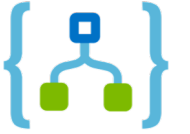


Connectors:

- FTP
- HTTP
- SQL Server
- Azure Blob Storage
- Office 365 Outlook
- Event Hub
- Service Bus
- Twitter
- Power BI
- Salesfroce
- Cognitive Services
- Dynamics 365 CRM / NAV
- Google Drive
- Youtube
- Informix
- DB2
- ...



Azure Logic App




Trigger:

Listener waiting for event A


e.g. new file created on a blob storage

Search: blob


Connectors [See more](#)

-  Azure Blob Storage

Triggers (1) [Actions \(9\)](#) [See more](#)

-  Azure Blob Storage - When one or more blobs are added or modified (metadata only) Preview [i](#)

TELL US WHAT YOU NEED

 Help us decide which connectors and triggers to add next with [UserVoice](#)



Azure Logic App



Action:

Follows after each trigger.

What to do when a trigger act.

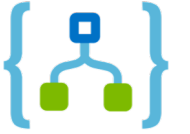
e.g. copy blob

The screenshot shows the search results for 'blob' in the Azure Logic App connector marketplace. The search bar at the top contains the text 'blob'. Below the search bar, there are two main sections: 'Connectors' and 'Triggers (1) Actions (9)'. The 'Connectors' section shows the 'Azure Blob Storage' connector icon and name. The 'Triggers (1) Actions (9)' section is currently selected, and it lists nine actions related to Azure Blob Storage, each with an information icon to its right. The actions listed are:

- Azure Blob Storage - Create blob
- Azure Blob Storage - Copy blob
- Azure Blob Storage - Delete blob
- Azure Blob Storage - Extract archive to folder
- Azure Blob Storage - Get blob content
- Azure Blob Storage - Get blob content using path
- Azure Blob Storage - Get Blob Metadata
- Azure Blob Storage - Get Blob Metadata using path



Azure Logic App



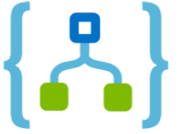
Integrationskonto-Connectors:

- A52
- EDIFACT
- XML
- X12



Azure Logic App

Demo ...



Azure Function



Azure Functions is a solution for easily running small pieces of code, or "functions," in the cloud. You can write just the code you need for the problem at hand, without worrying about a whole application or the infrastructure to run it.



Azure Function



Language:

- C#
- F#
- Node.js
- Python
- PHP
- Batch
- Bash
- any executable



Azure Function



Integrations:

- Azure Cosmos DB
- Azure Event Hubs
- Azure Mobile Apps (tables)
- Azure Notification Hubs
- Azure Service Bus (queues and topics)
- Azure Storage (blob, queues, and tables)
- GitHub (webhooks)
- On-premises (using Service Bus)
- Twilio (SMS messages)



Azure Function



What can I do:

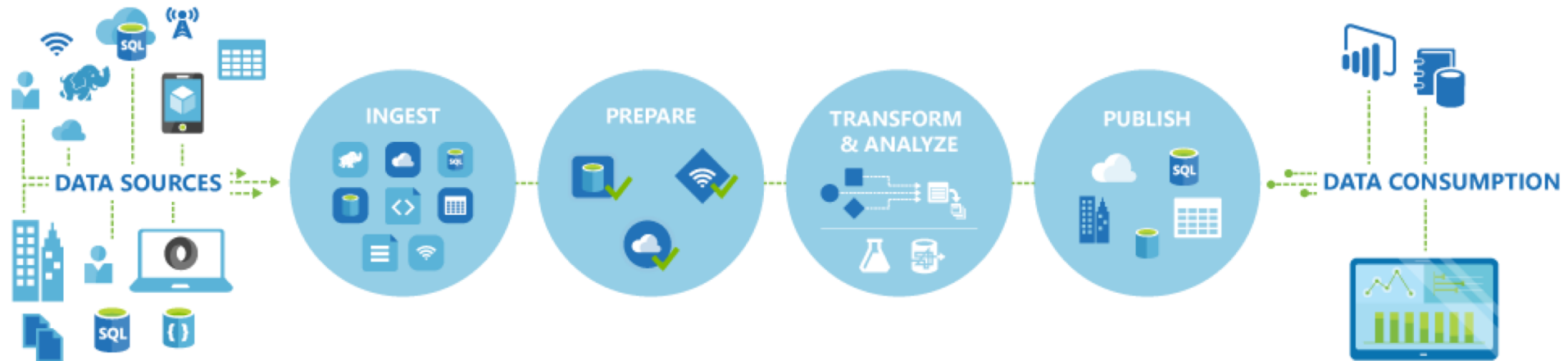
- BlobTrigger
- EventHubTrigger
- Generic webhook
- GitHub webhook
- HTTPTrigger
- QueueTrigger
- ServiceBusQueueTrigger
- ServiceBusTopicTrigger
- TimerTrigger



Azure Data Factory (ADF)



Cloud-based data integration service that allows you to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation.



Azure Data Factory (ADF)



Source:

- Azure Storage
- FTP
- HTTP
- Amazon S3
- HDFS
- Oracle
- SAP BW
- SAP HANA

Sink:

- Azure Blob Storage
- Azure Data Lake
- Azure SQL DB
- Azure SQL DW
- Azure Cosmos DB
- Oracle
- Filesystem



Azure Data Factory (ADF)



Pipeline:

A data factory may have one or more pipelines. A pipeline is a group of activities. Together, the activities in a pipeline perform a task.



Azure Data Factory (ADF)



Activity:

Activities define the actions to perform on your data. For example, you may use a Copy activity to copy data from one data store to another data store.



Azure Data Factory (ADF)



Datasets:

An activity takes zero or more datasets as inputs and one or more datasets as outputs. Datasets represent data structures within the data stores, which simply point or reference the data you want to use in your activities as inputs or outputs.



Azure Data Factory (ADF)

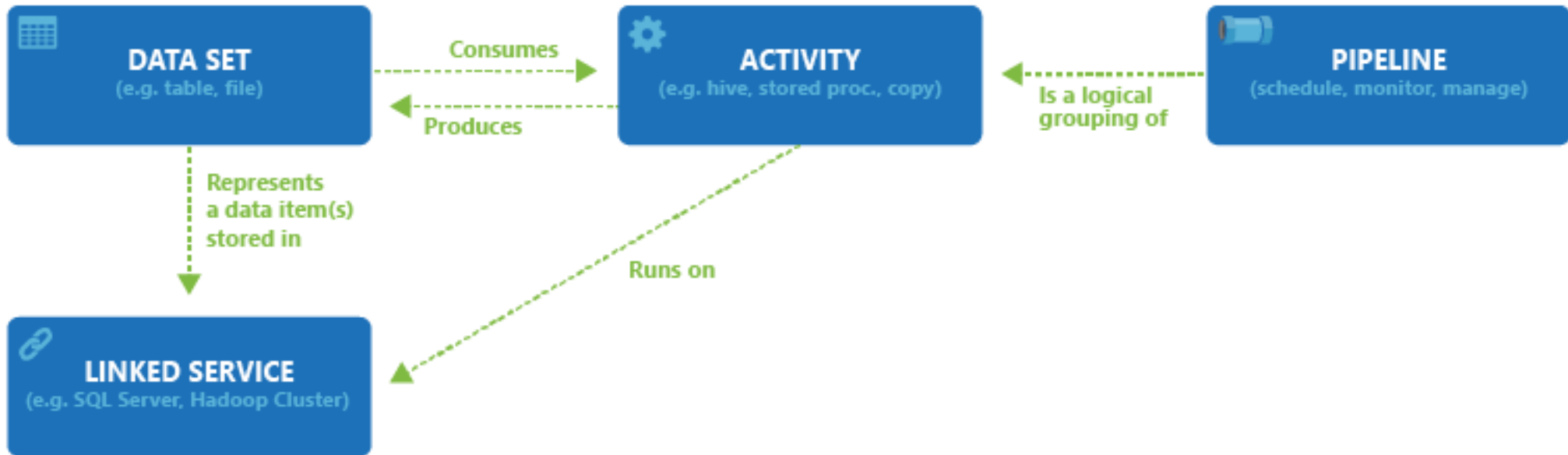


Linked services:

Linked services are much like connection strings, which define the connection information needed for Data Factory to connect to external resources. Think of it this way - a linked service defines the connection to the data source and a dataset represents the structure of the data.



Azure Data Factory (ADF)



Azure Data Factory (ADF)



Data transformation activity	Compute environment
Hive	HDInsight [Hadoop]
Pig	HDInsight [Hadoop]
MapReduce	HDInsight [Hadoop]
Hadoop Streaming	HDInsight [Hadoop]
Spark	HDInsight [Hadoop]
Machine Learning activities: Batch Execution and Update Resource	Azure VM
Stored Procedure	Azure SQL, Azure SQL Data Warehouse, or SQL Server
Data Lake Analytics U-SQL	Azure Data Lake Analytics
DotNet	HDInsight [Hadoop] or Azure Batch



Azure Data Factory (ADF)



Demo ...



Azure Data Factory version 2



Azure Data Factory version 2 builds upon the original Azure Data Factory data movement and transformation service, extending to a broader set of cloud-first data integration scenarios. Azure Data Factory Version 2 brings the following capabilities:

- **Control Flow and Scale**
- **Deploy and run SSIS packages in Azure**

With version 2, you can also migrate existing SQL Server Integration Services (SSIS) packages to the cloud to lift & shift SSIS as an Azure service managed within ADF utilizing a new feature of “Integration Runtimes” (IR). By spinning-up an SSIS IR in version 2, you have the ability to execute, manage, monitor, and build SSIS packages in the cloud.



Integration runtime in Azure Data Factory V2

The Integration Runtime (IR) is the compute infrastructure used by Azure Data Factory to provide the following data integration capabilities across different network environments:

- **Data movement:** Move data between data stores in public network and data stores in private network (on-premise or virtual private network). It provides support for built-in connectors, format conversion, column mapping, and performant and scalable data transfer.
- **Activity dispatch:** Dispatch and monitor transformation activities running on a variety of compute services such as Azure HDInsight, Azure Machine Learning, Azure SQL Database, SQL Server, and more.
- **SSIS package execution:** Natively execute SQL Server Integration Services (SSIS) packages in a managed Azure compute environment.



Integration runtime in Azure Data Factory V2



Web Based Developer UI

The screenshot displays the Azure Data Factory web-based developer UI. The main interface is divided into several sections:

- Left Panel:** A navigation pane showing the hierarchy of the 'CustomerChurnFactory' workspace, including Pipelines, Data Flows, Datasets, and Integration Runtimes.
- Activities Panel:** A list of available activities such as AzureML, Custom.Net, Data Prep, Hive, Map Reduce, Pig, Stored Procedure, Spark, Data Flow, and HTTP.
- Design Canvas:** A central workspace where a pipeline is being designed. It shows a 'Data Flow' activity connected to a 'Spark' activity, which is then connected to an 'HTTP' activity (SendEmail) and another 'Data Flow' activity (CustomerChurned). Below the canvas is an 'Output Log' section showing a message: '7/12/2017 8am UTC: Started CallDataRecords & CustomerInfo to CustomerChurned activity.'
- Toolbox:** A panel on the right side of the canvas containing 'Source' and 'Sink' categories with various connectors like Azure Blob Storage, Amazon S3, Azure SQL Database, and Azure SQL Data Warehouse.
- Settings Panel:** A panel on the far right showing the 'Mapping' settings for a data flow. It includes a 'Mapping Options' dropdown set to 'Automatic', an 'Auto Map' button, and a table showing the mapping between source and sink fields.

Source fields: 25 / 25 mapped		Sink fields: 25 / 25 mapped	
FIELD	TYPE	FIELD	TYPE
Age	int	Age	int
AnnualIncome	BigInt	AnnualIncome	BigInt
CallDropRate	Double	CallDropRate	Double
CallFailureRate	Double	CallFailureRate	Double
CallingNum	String	CallingNum	String
CustomerID	Int	CustomerID	Int
CustomerSuspended	String	CustomerSuspended	String
Education	String	Education	String
Gender	String	Gender	String
HomeOwner	String	HomeOwner	String
MaritalStatus	String	MaritalStatus	String
MonthlyBilledAmount	Int	MonthlyBilledAmount	Int
NoAdditionalLines	Int	NoAdditionalLines	Int

Azure Data Lake (ADL)



Azure Data Lake Store

Azure Data Lake Analytics

HDFS for the Cloud

Hadoop

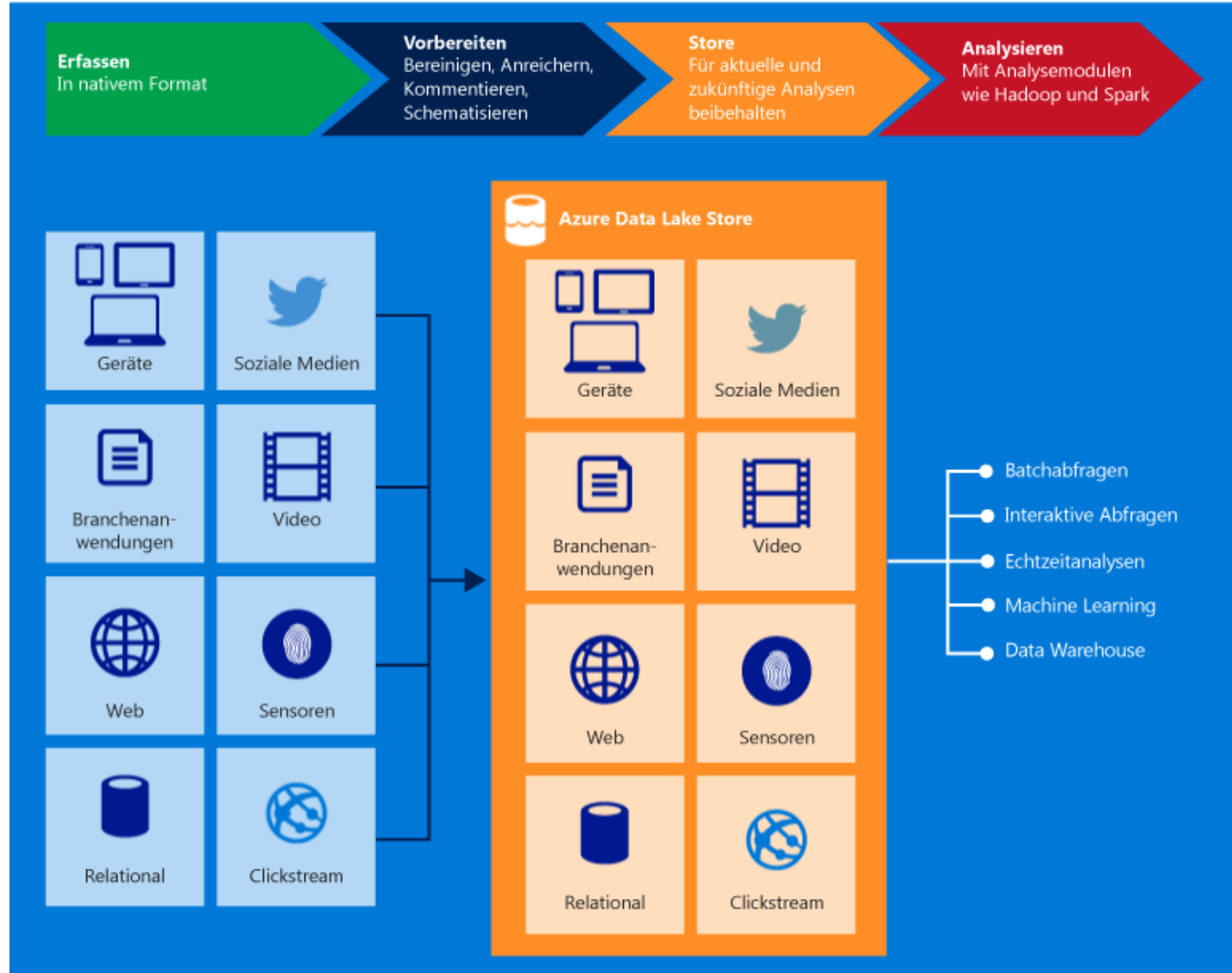
Spark

Always encrypted

Big Data



Azure Data Lake (ADL)



Azure Data Lake (ADL)

U-SQL

Dynamic scaling

HDFS for the Cloud



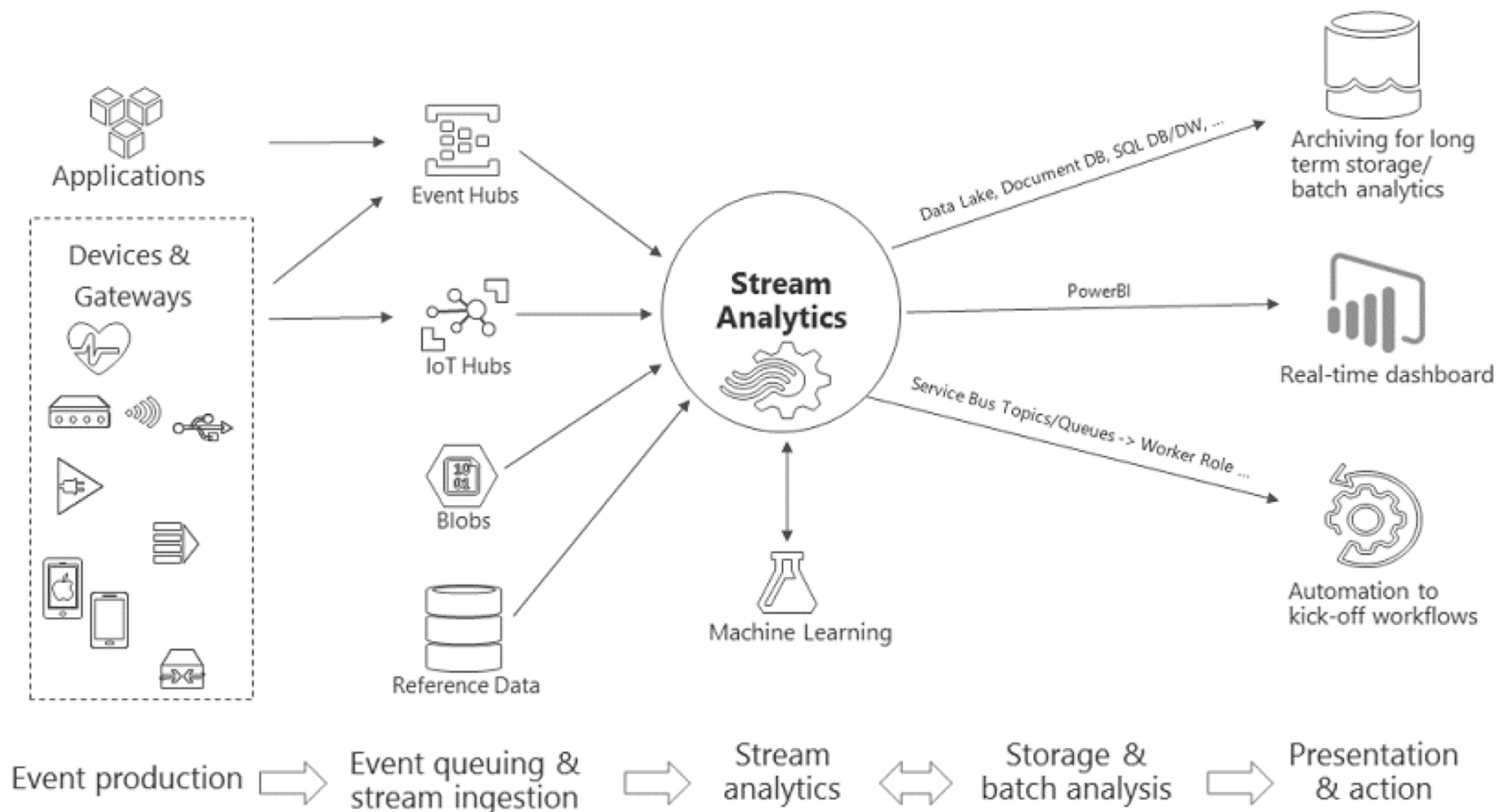
Azure Stream Analytics (ASA)



Azure Stream Analytics is a fully managed event-processing engine that lets you set up real-time analytic computations on streaming data. The data can come from devices, sensors, web sites, social media feeds, applications, infrastructure systems, and more.



Azure Stream Analytics (ASA)



Event production ⇒ Event queuing & stream ingestion ⇒ Stream analytics ⇔ Storage & batch analysis ⇒ Presentation & action

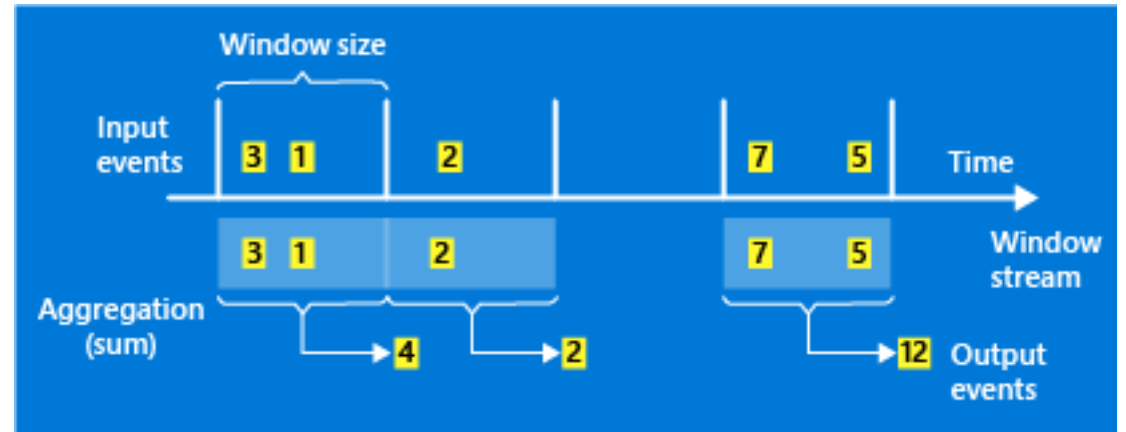


Azure Stream Analytics (ASA)



Grouping:

- Tumbling Window
- Hopping Window
- Sliding Window



Azure Stream Analytics (ASA)



Source:

- Azure Event Hub
- Azure IoT Hub
- Azure Blob Storage

Supported formats:

- Avro
- JSON
- CSV

Sink:

- Azure Blob Storage
- Azure Data Lake Store
- Azure Document DB
- Azure Event Hub
- Azure Table Storage
- Azure SQL DB
- Azure Service Bus Queue
- Power BI



Azure Automation



Azure Automation is a software as a service (SaaS) application that provides a scalable and reliable, multi-tenant environment to automate processes with runbooks and manage configuration changes to Windows and Linux systems using Desired State Configuration (DSC) in Azure, other cloud services, or on-premises.



Azure Runbook



Types:

graphical runbook

PowerShell runbook

Resources

Solutions	Runbooks 37	Jobs	64 ASSETS	Hybrid Worker Groups 6
DSC Configurations 0	DSC Nodes 0			

The screenshot shows the Azure Runbook editor interface. On the left is a 'Library' pane with a tree view of activity categories: CMDLETS (Azure, GitHub, Microsoft.PowerShell.Core, Microsoft.PowerShell.Diagnostics, Microsoft.PowerShell.Management, Microsoft.PowerShell.Security, Microsoft.PowerShell.Utility, Microsoft.WSMan.Management, Twilio, Twitter), RUNBOOKS (All), ASSETS (Variables, Connections, Credentials, Certificates), and RUNBOOK CONTROL (Junction, Workflow Script). The main area is a 'Canvas' for building a runbook, with a flowchart showing a sequence of activities. The right pane is for 'Configuration' of the selected activity. The top toolbar includes Save, Publish, Revert to published, Input and output, and Test pane.



Visual Studio & TFS

- Azure Data Factory
- Azure Data Lake
- Azure Function
- Azure Logic App *
- Azure Stream Analytics (not all sources and destinations supported !)



Deployment options

	Azure Portal	Visual Studio	PowerShell
Azure Data Factory	X	X	X
Azure Data Lake	X	X	X
Azure Function	X	X	X
Azure Logic App	X	(X)	X
Azure Stream Analytics	X	(X)	X



Always keep in mind

- Error handling
- Notification
- Reporting
- Data delivery



Question! Question?

Thank you for your attention

✉ a.klein@consulting-bi.de

🐦 @SQL_Alex

🏠 consulting-bi.de

