Azure as a Big Data Platform

Tomasz Krawczyk

tkrawczyk@future-processing.com

Agenda

- Big Data
  - Processing (Theory)
- First challenge
- Azure Platform
  - Big Data Services
- Demo and Q&A

3Vs of Big Data

40 Zetta bytes by 2020 and 163 Zetta bytes by 2025

› Data Volume

Byte	One grain of rice
Kilobyte	Cup of rice
Megabyte	8 bags of rice
Gigabyte	3 semi trucks
Terabyte	2 container ships
Petabyte	Blankets Manhattan
Exabyte	Blankets west coast states
Zettabyte	Fills the Pacific Ocean
Yottabyte	As earth-sized rice ball

› Data Variety

- Structured
- Unstructured
- Semi-structured
- All the above

› Data Velocity

- **Near to Real Time**
- **Batch**

Schema-on-Read vs Schema-on-Write

SCHEMA-ON-READ (HADOOP OR ADLS):

- Copy data in its native format
- Create schema + parser
- Query Data in its native format (does ETL on the fly)

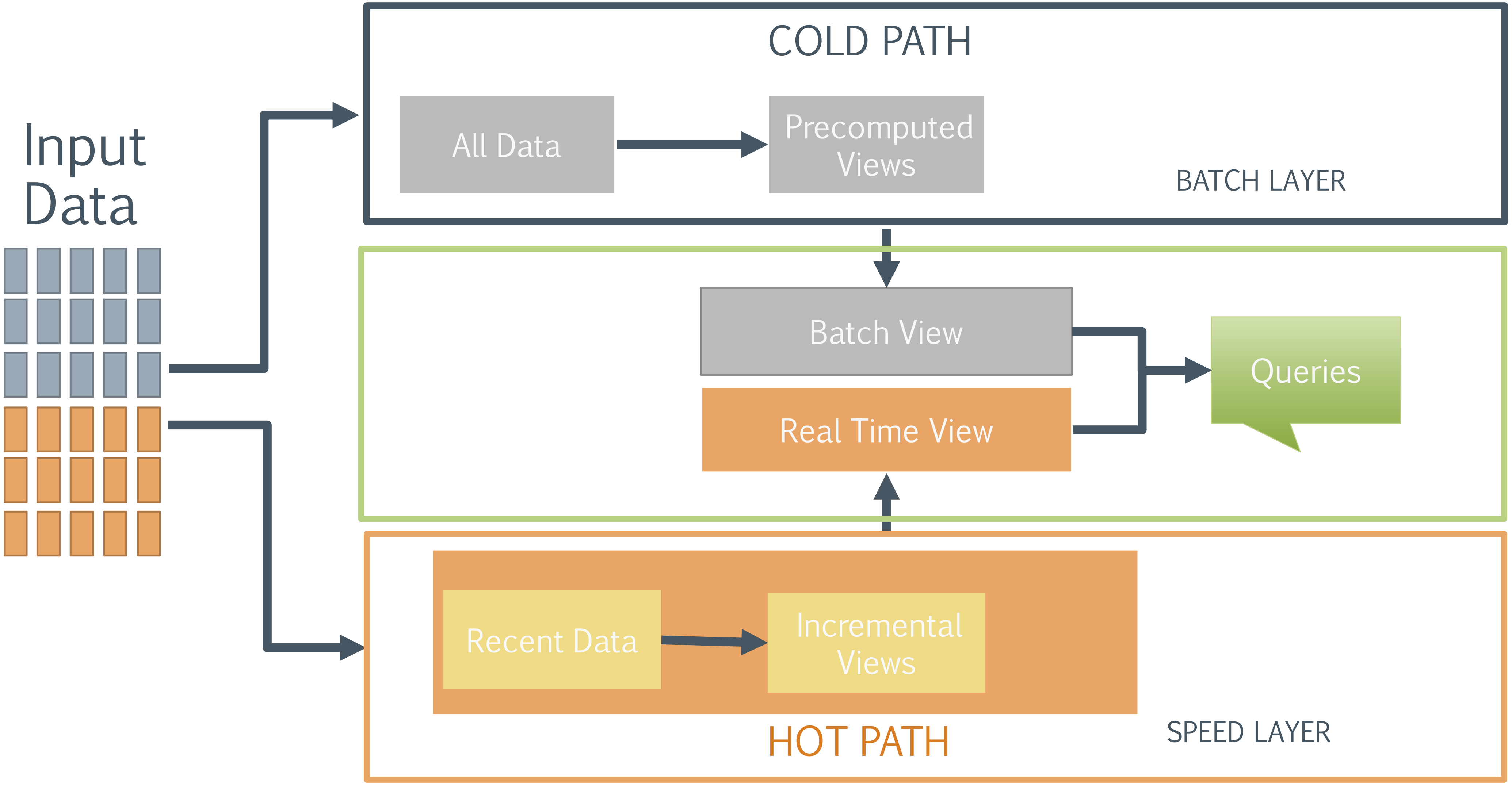
New data can start flowing any time and will appear retroactively once the schema/parser properly describes it.

SCHEMA-ON-WRITE (RDBMS):

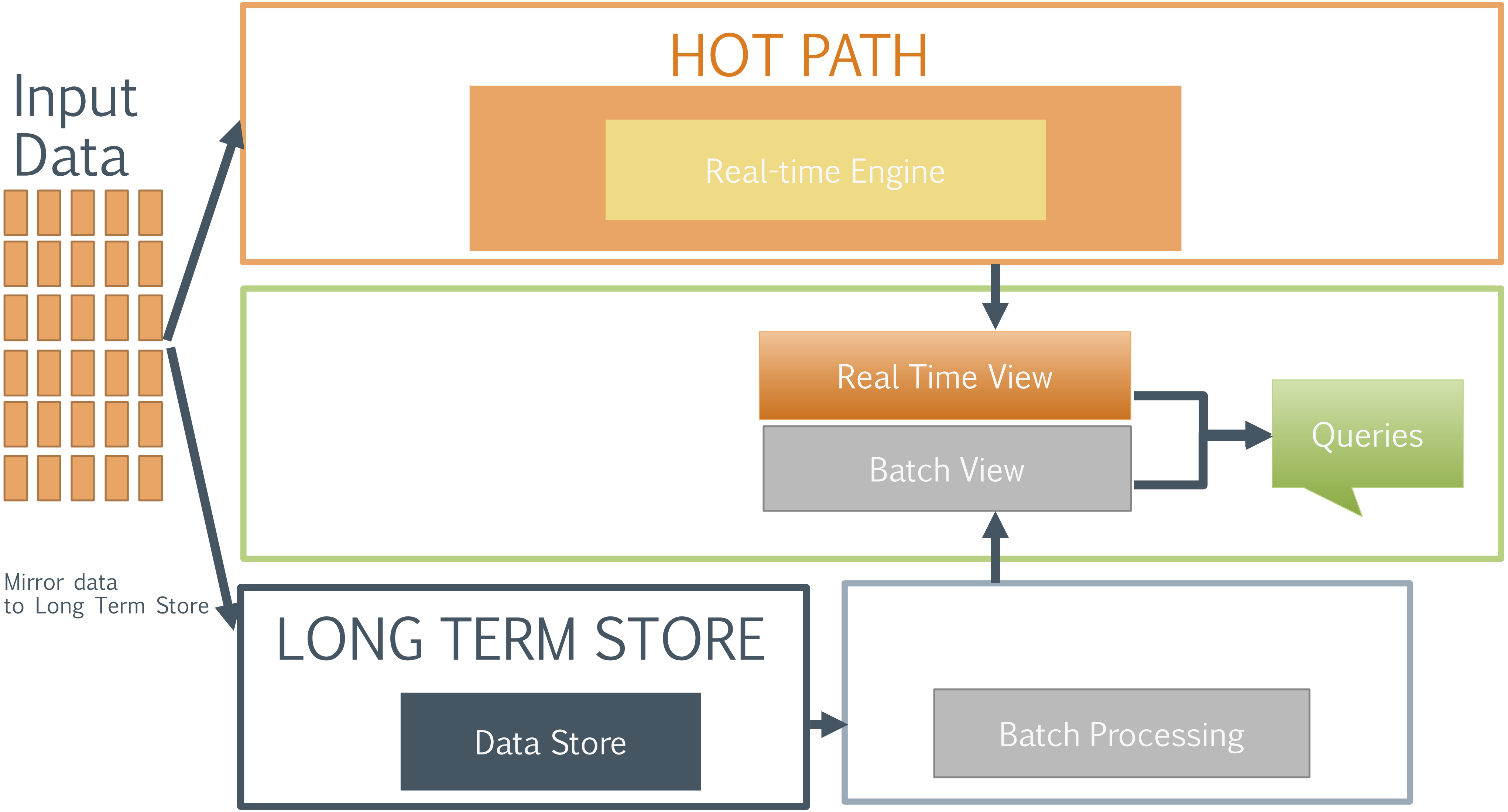
- Create static DB schema
- Transform data into RDBMS
- Query data in RDBMS format

New columns must be added explicitly before new data can propagate into the system.

Lambda architecture



Kappa architecture



Data Lake Approach

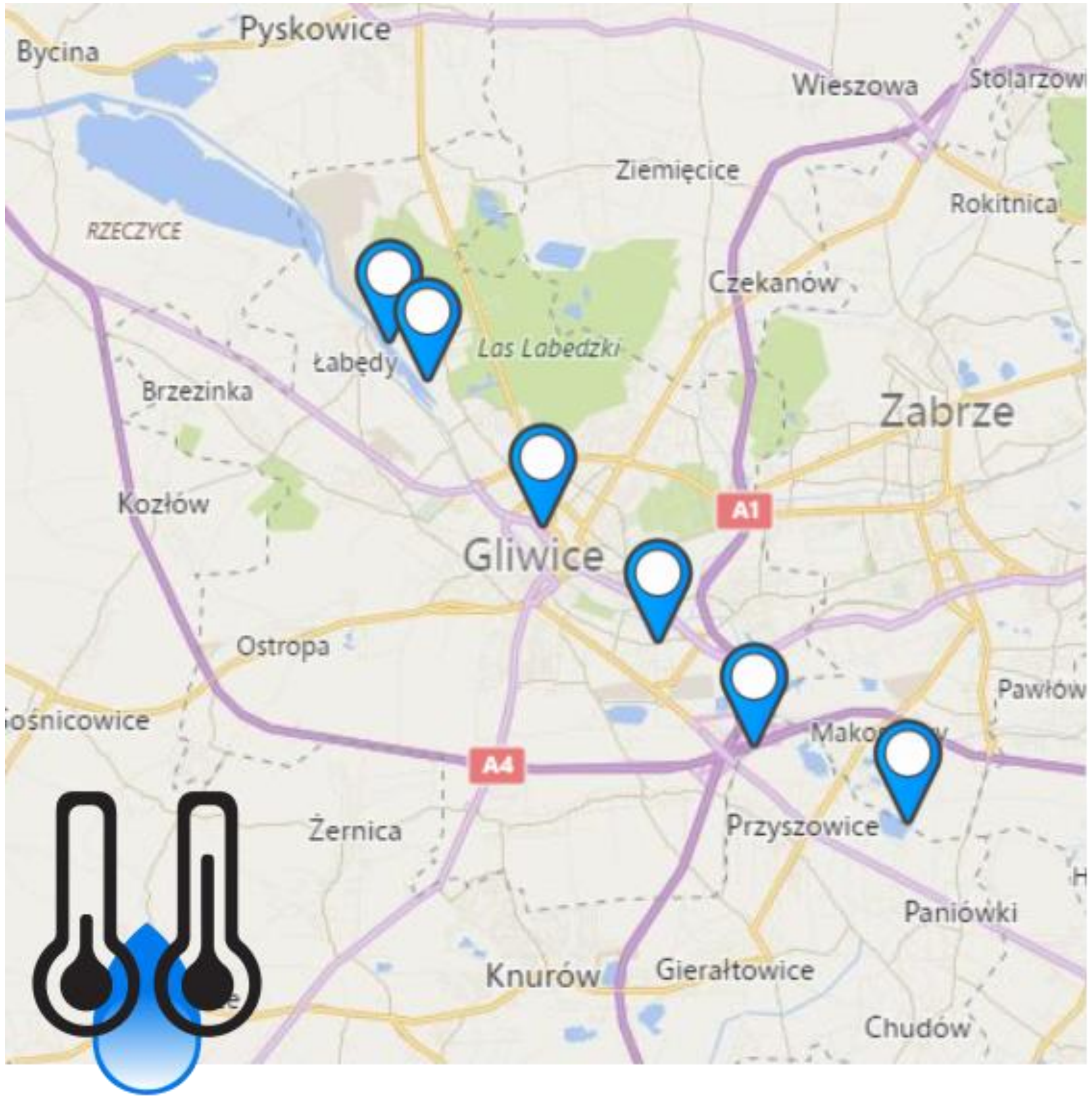
› What is a Data Lake ?

„If you think of a **datamart** (asubset of a data warehouse) as a store of bottled water – cleansed and packaged and structured for easy consumption – the **data lake** is a large body of water in a more **natural state** „

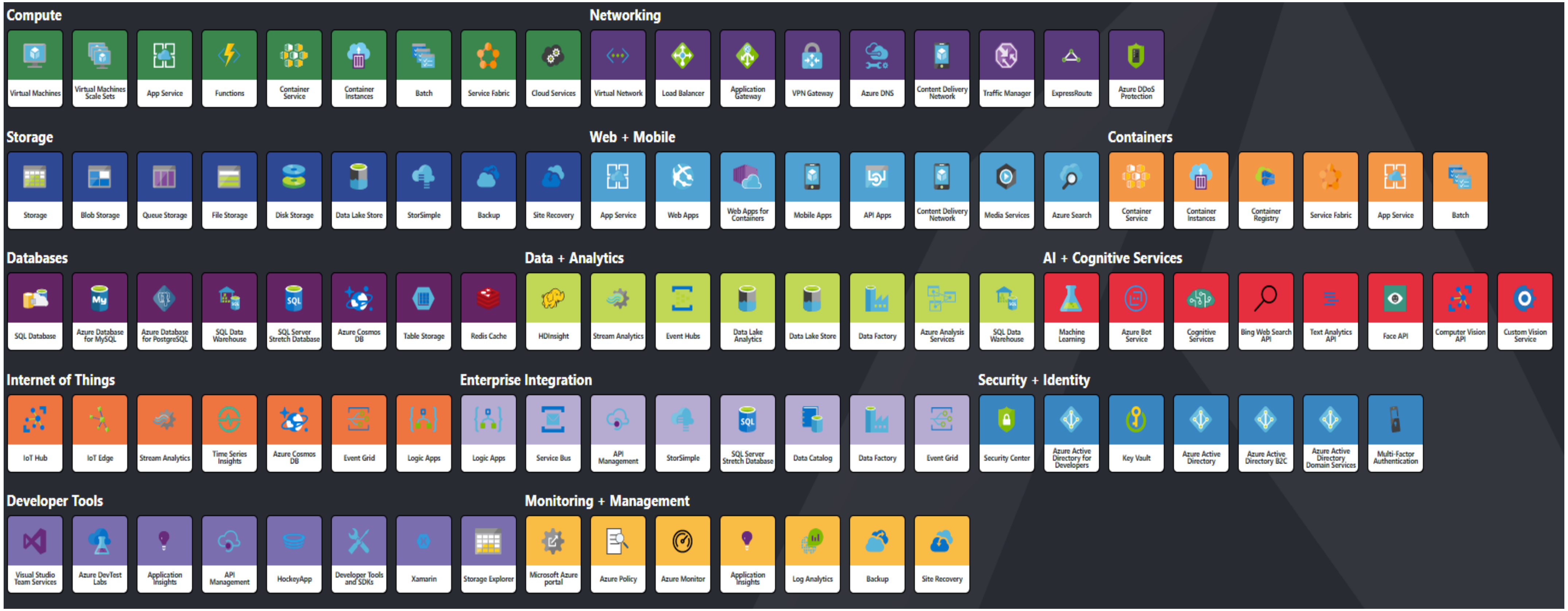
Pentaho CTO James Dixon

First Challange

- PoC
  - Measuring devices
  - Online monitoring
  - Daily Statistics
  - Historical Data



Azure Services



Compute

Virtual Machines
Virtual Machines Scale Sets
App Service
Functions
Container Service
Container Instances
Batch
Service Fabric
Cloud Services

Networking

Virtual Network
Load Balancer
Application Gateway
VPN Gateway
Azure DNS
Content Delivery Network
Traffic Manager
ExpressRoute
Azure DDoS Protection

Storage

Storage
Blob Storage
Queue Storage
File Storage
Disk Storage
Data Lake Store
StorSimple
Backup
Site Recovery

Web + Mobile

App Service
Web Apps
Web Apps for Containers
Mobile Apps
API Apps
Content Delivery Network
Media Services
Azure Search

Containers

Container Service
Container Instances
Container Registry
Service Fabric
App Service
Batch

Databases

SQL Database
Azure Database for MySQL
Azure Database for PostgreSQL
SQL Data Warehouse
SQL Server Stretch Database
Azure Cosmos DB
Table Storage
Redis Cache

Data + Analytics

HDInsight
Stream Analytics
Event Hubs
Data Lake Analytics
Data Lake Store
Data Factory
Azure Analysis Services
SQL Data Warehouse

AI + Cognitive Services

Machine Learning
Azure Bot Service
Cognitive Services
Bing Web Search API
Text Analytics API
Face API
Computer Vision API
Custom Vision Service

Internet of Things

IoT Hub
IoT Edge
Stream Analytics
Time Series Insights
Azure Cosmos DB
Event Grid
Logic Apps

Enterprise Integration

Logic Apps
Service Bus
API Management
StorSimple
SQL Server Stretch Database
Data Catalog
Data Factory
Event Grid

Security + Identity

Security Center
Azure Active Directory for Developers
Key Vault
Azure Active Directory
Azure Active Directory B2C
Azure Active Directory Domain Services
Multi-Factor Authentication

Developer Tools

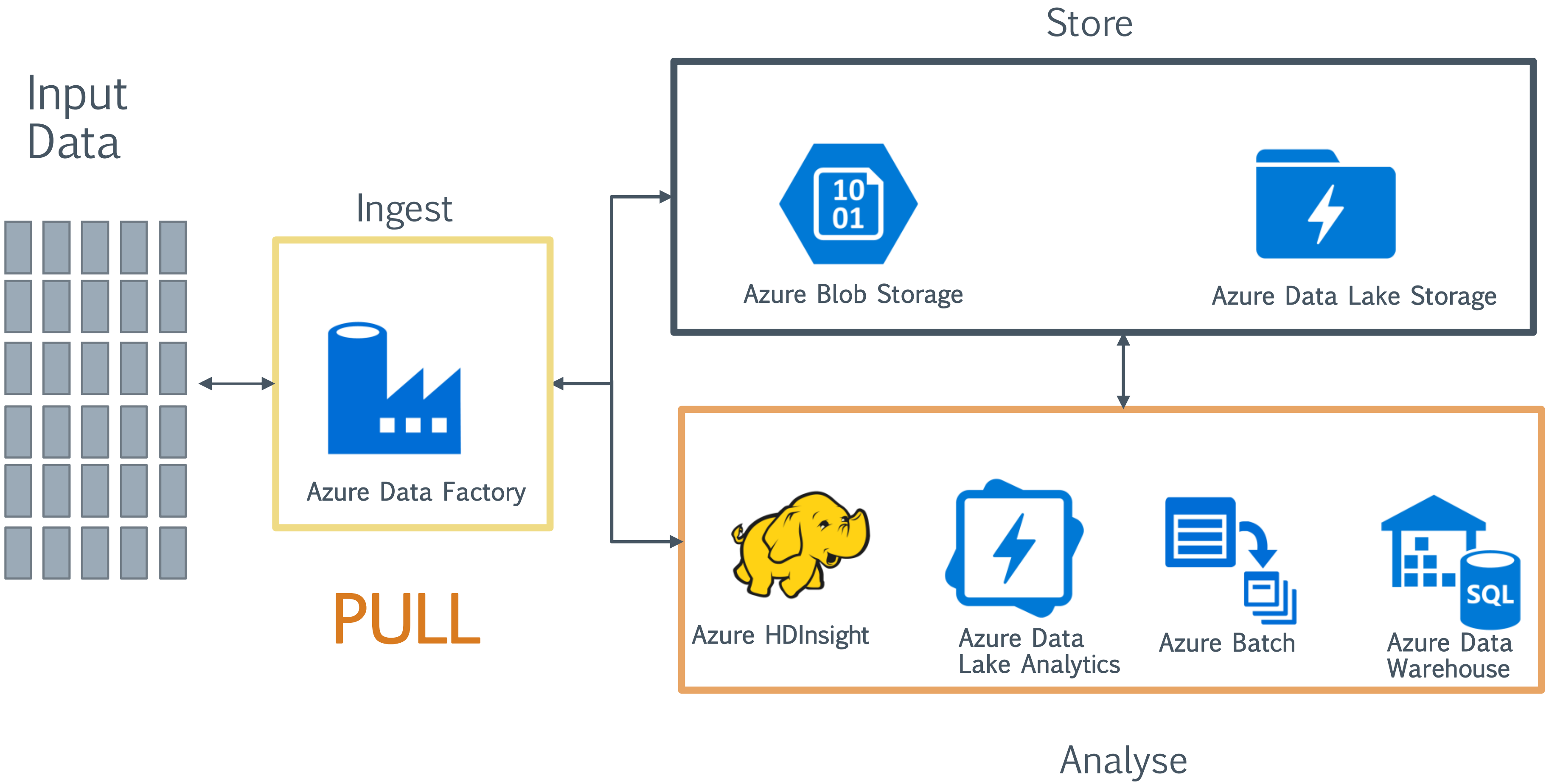
Visual Studio Team Services
Azure DevTest Labs
Application Insights
API Management
HockeyApp
Developer Tools and SDKs
Xamarin
Storage Explorer

Monitoring + Management

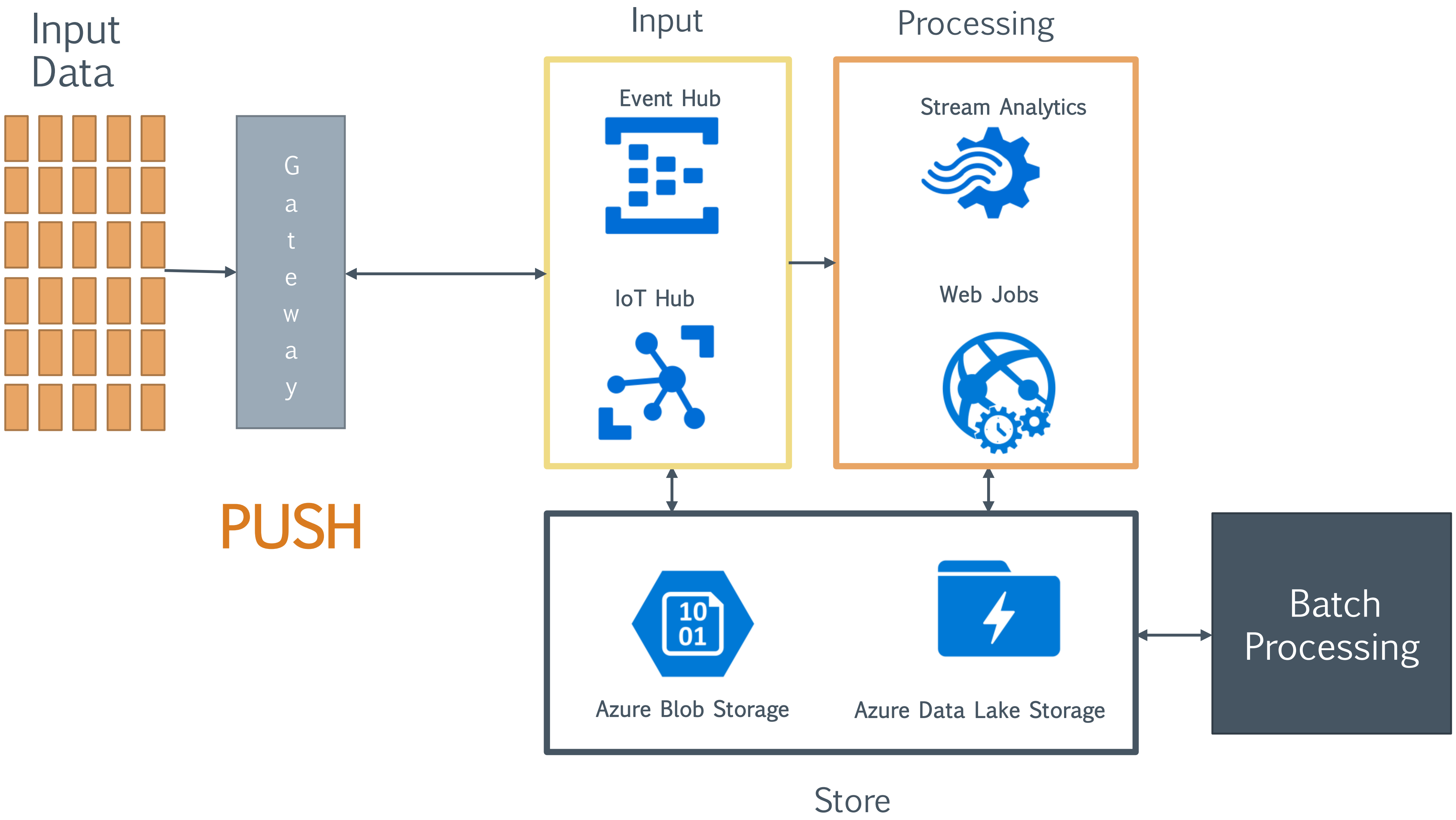
Microsoft Azure portal
Azure Policy
Azure Monitor
Application Insights
Log Analytics
Backup
Site Recovery

http://azureinteractives.azurewebsites.net/Azure101Cards/default.html

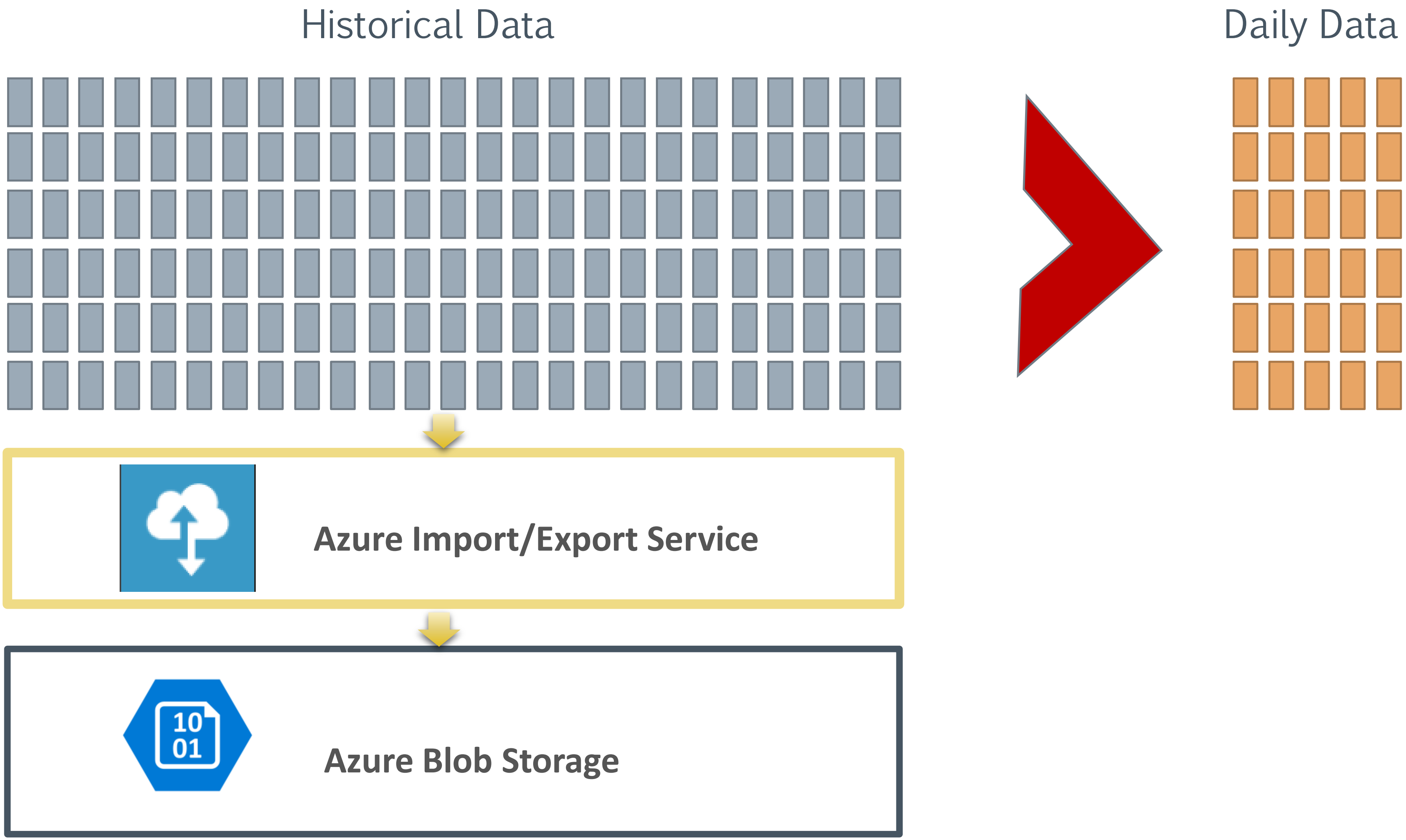
Azure – Lambda architecture (Cold Path)



Azure –Kappa architecture



Azure Initial Load



Historical Data

Daily Data

Azure Import/Export Service

Azure Blob Storage

Data Lake Store vs Blob Storage

	Azure Data Lake Store	Azure Blob Storage
Purpose	Optimized storage for big data analytics workloads	General purpose object store for a wide variety of storage scenarios
Use Cases	Batch, interactive, streaming analytics and machine learning data such as log files, IoT data, click streams, large datasets	Any type of text or binary data, such as application back end, backup data, media storage for streaming and general purpose data
Key Concepts	Data Lake Store account contains folders, which in turn contains data stored as files	Storage account has containers, which in turn has data in the form of blobs
Structure	Hierarchical file system	Object store with flat namespace
API	REST API over HTTPS	REST API over HTTP/HTTPS
Hadoop File System Client	Yes	Yes
Data Operations - Authentication	Based on Azure Active Directory Identities	Based on shared secrets - Account Access Keys and Shared Access Signature Keys.
Data Operations - Authorization	POSIX Access Control Lists (ACLs). ACLs based on Azure Active Directory Identities can be set file and folder level.	For account-level authorization – Use Account Access Keys For account, container, or blob authorization - Use Shared Access Signature Keys

Azure Event Hub vs IoT Hub

- **Azure Event Hub** is a highly scalable data streaming platform and event ingestion service, capable of receiving and processing millions of events per second
- **Azure IoT Hub** is a fully managed service that enables reliable and secure bidirectional communications between millions of IoT devices and a solution back end

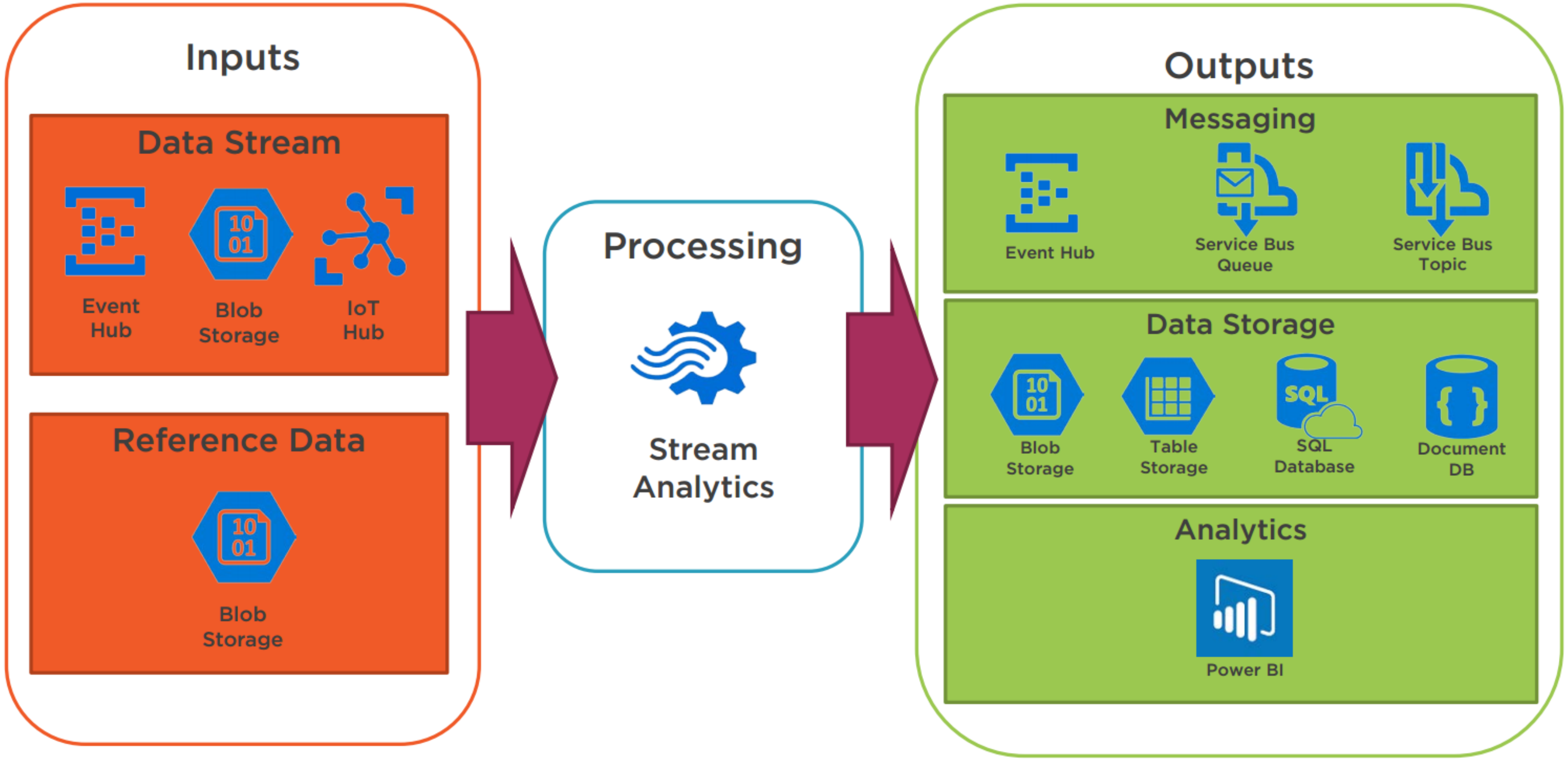
Area	IoT Hub	Event Hub
Device protocol support	Supports MQTT, MQTT over WebSockets, AMQP, AMQP over WebSockets, and HTTPS.	Supports AMQP, AMQP over WebSockets, and HTTPS.
Device state information	Device twins can store and query device state information.	No device state information can be stored.

Source:https://docs.microsoft.com/en-us/azure/iot-hub/iot-hub-compare-event-hubs

Azure Stream Analytics

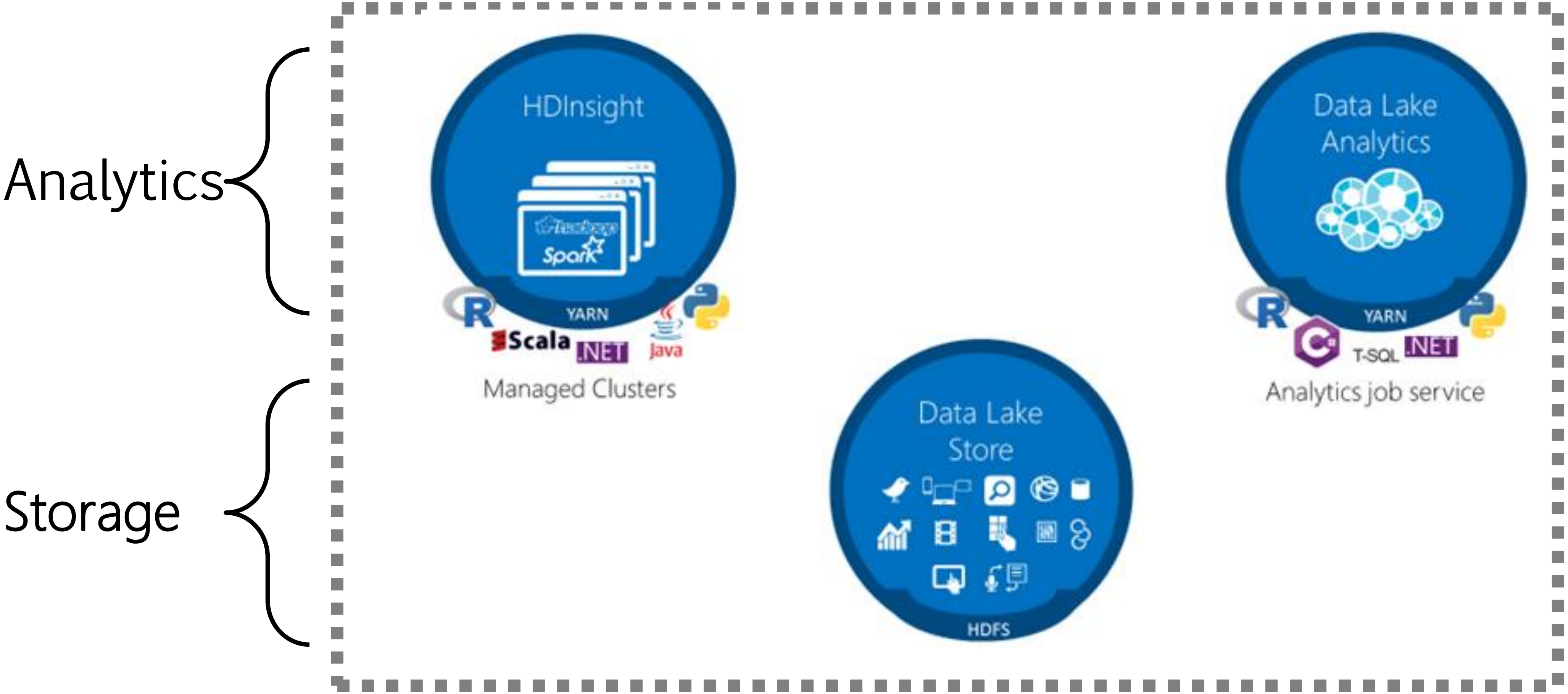
- Azure Stream Analytics is a managed event-processing engine set up real-time analytic computations on streaming data.
- Data Sources
  - **Data Stream**
  - **Reference Data**
- SQL-like language for querying live data streams
  - Supports SELECT, FROM, WHERE, GROUP BY, and other common Data Manipulation Language (DML) statements
  - Supports COUNT, AVG, DATEDIFF, and other common functions
- Supports temporal grouping of events via "windowing“
  - Tumbling Window, Hopping Window, **Sliding** Window

Azure Stream Analytics at work



Source: Alan Smith

Data Lake approach on Azure



Analytics

Storage

Azure HDInsight vs Azure Data Lake Analytics

- Azure HDInsight
  - Cluster as a Service
  - Hadoop, Hbase,Storm, Spark, R Server, Kafka

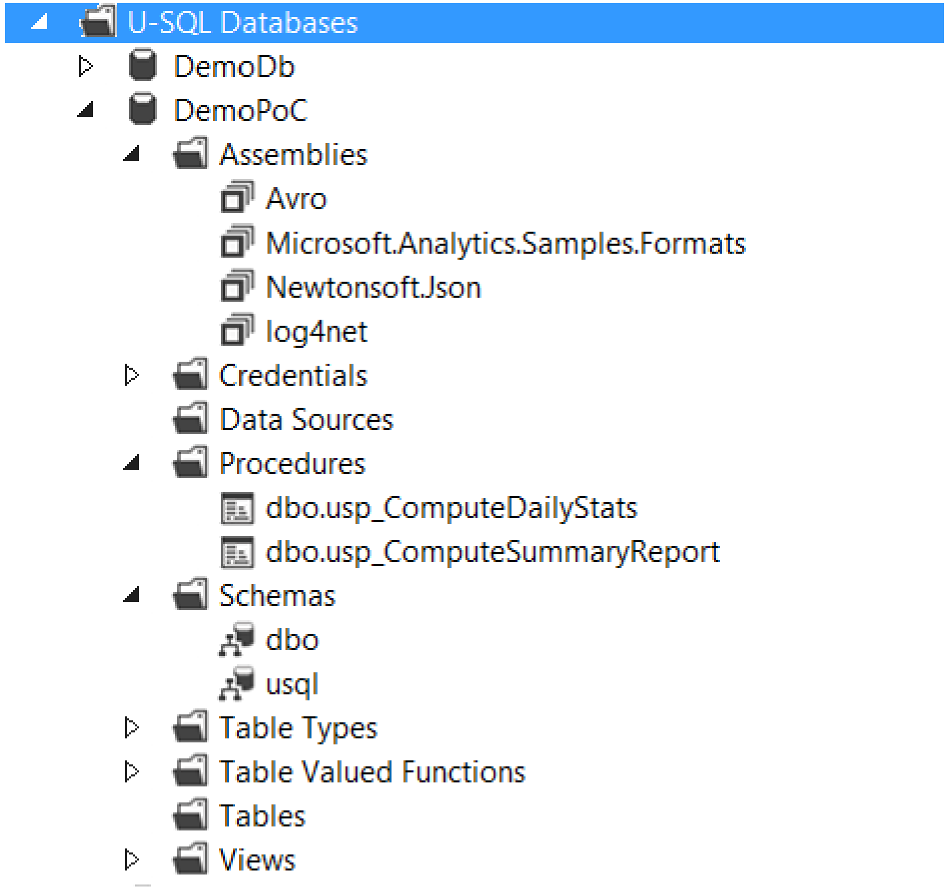
- Azure Data Lake Analytics
  - Job/Query as a Service
  - USQL (.Net ,Python, R Language, Cognitive)

Azure Data Lake Analytics

- **A distributed analytics service built on Apache YARN that dynamically scales to your needs**
  - Pay PER QUERY & Scale PER QUERY
  - FEDERATED QUERY across Azure data sources
  - Includes U-SQL, a language that unifies the benefits of SQL with the expressive power of C#
  - No limits to SCALE
  - Optimized to work with ADL STORE

U–SQL - A new language for Big Data

```
DECLARE @projectsInput string = @"Projects\{file}.csv";
DECLARE @eventDate DateTime =
System.DateTime.Parse("2017/04/01");
DECLARE @numbers int = 2;
REFERENCE ASSEMBLY USQLCSharpDemo;
USING ImageColorsProcessor =
USQLCSharpDemo.ImageColorProducer;
@projects =
    EXTRACT project string,
            startDate DateTime,
            endDate DateTime,
            file string
    FROM @projectsInput
    USING Extractors.Csv(skipFirstNRows : 1, quoting :
true);
@agg =
    SELECT project,
           COUNT( * ) AS units
    FROM @details WHERE project.StartsWith("My")
    GROUP BY project;
@myprojects =
    SELECT us.project,
           p.endDate
    FROM @details AS us
         JOIN
             @projects AS p
         ON p.project == us.project
    WHERE user.StartsWith("Me")
ORDER BY p.endDate DESC
FETCH 10 ROWS;
OUTPUT @myprojects
TO "myprojects.csv"
USING Outputters.Csv();
```



- U-SQL Databases
  - DemoDb
  - DemoPoC
    - Assemblies
      - Avro
      - Microsoft.Analytics.Samples.Formats
      - Newtonsoft.Json
      - log4net
    - Credentials
    - Data Sources
    - Procedures
      - dbo.usp_ComputeDailyStats
      - dbo.usp_ComputeSummaryReport
    - Schemas
      - dbo
      - usql
    - Table Types
    - Table Valued Functions
    - Tables
    - Views

USQL (+ .Net ,Python, R Language, Cognitive)

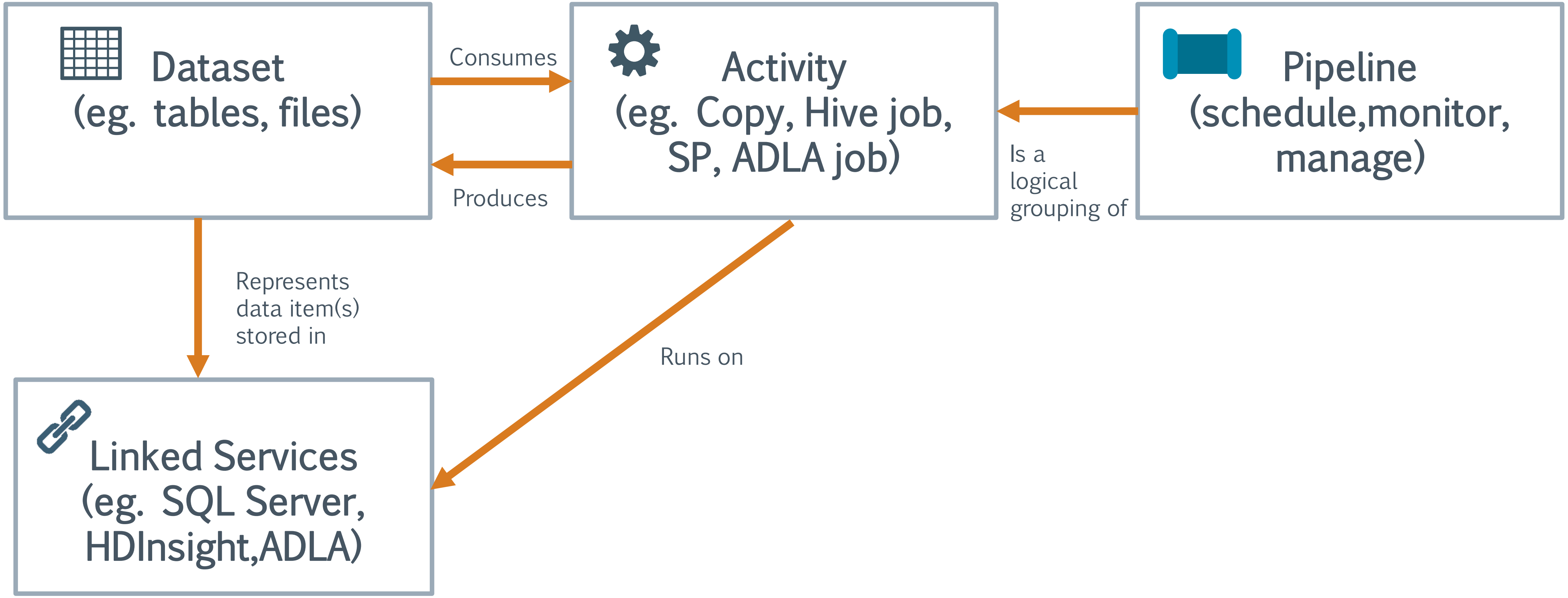
Azure Data Factory

- Fully managed service to support orchestration of data movement and transformation
- Connect to relational or non-relational data that is on-premises or in the cloud
- Allows monitor and manage data processing pipelines
- Version 1 and 2 (+SSIS)

Azure Data Factory Versions

- Azure Data Factory V1 - **GA Version**
- Azure Data Factory V2 - Public Preview+Designer (2018-01-16)
- What’s new
  - New pipeline model
    - **Rich pipeline orchestration**
    - **Triggers –ondemand,schedule,events**
  - SSIS Package Execution
    - Lift my existing packages to the cloud
  - Author & Monitor
    - Python,.Net
    - Visual Tools
  - Data Movement as as Service
    - Cloud,Hybrid
    - 64 connectors

Azure Data Factory V1 Pipelines



Dataset
(eg. tables, files)

Consumes

Produces

Activity
(eg. Copy, Hive job, SP, ADLA job)

Is a logical grouping of

Pipeline
(schedule,monitor, manage)

Represents data item(s) stored in

Runs on

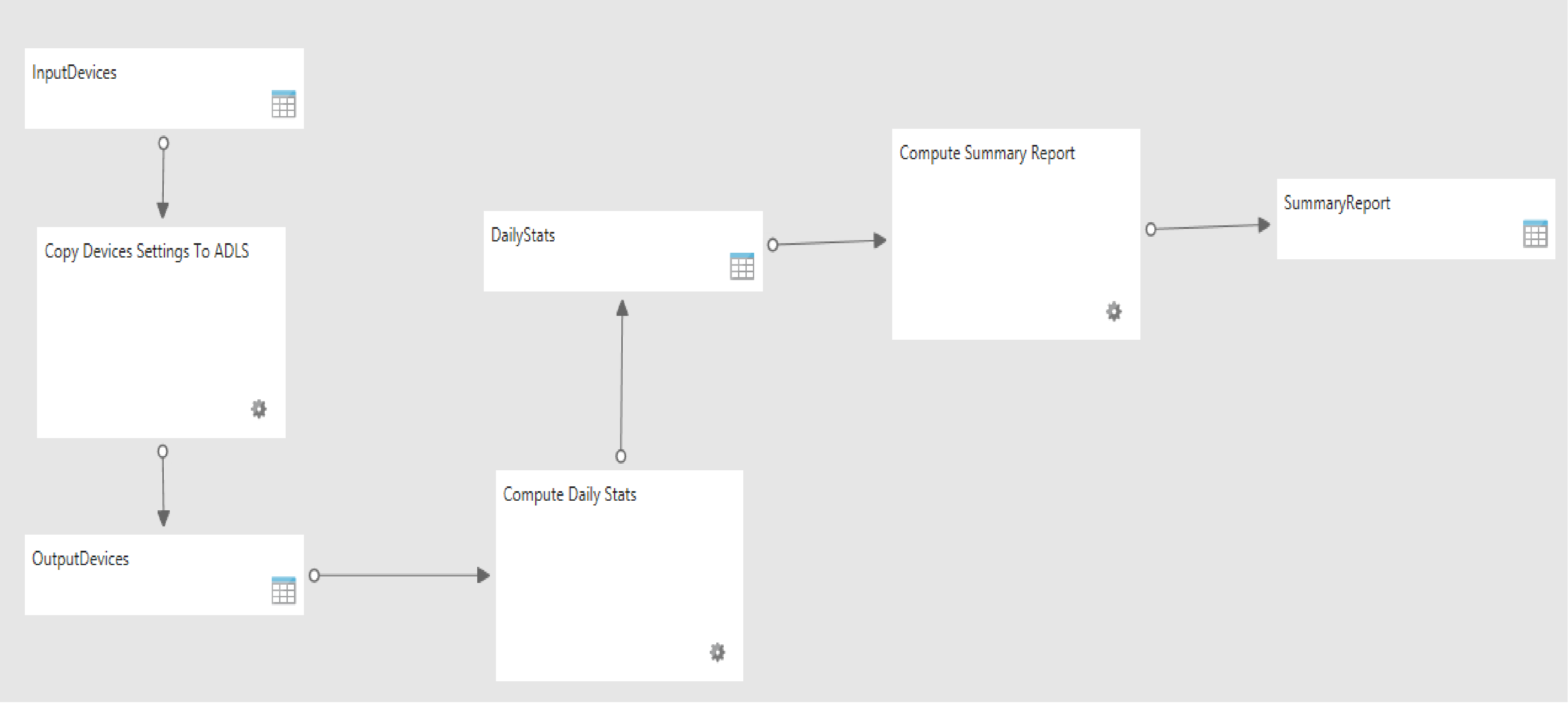
Linked Services
(eg. SQL Server, HDInsight,ADLA)

Azure Data Factory
Data activities

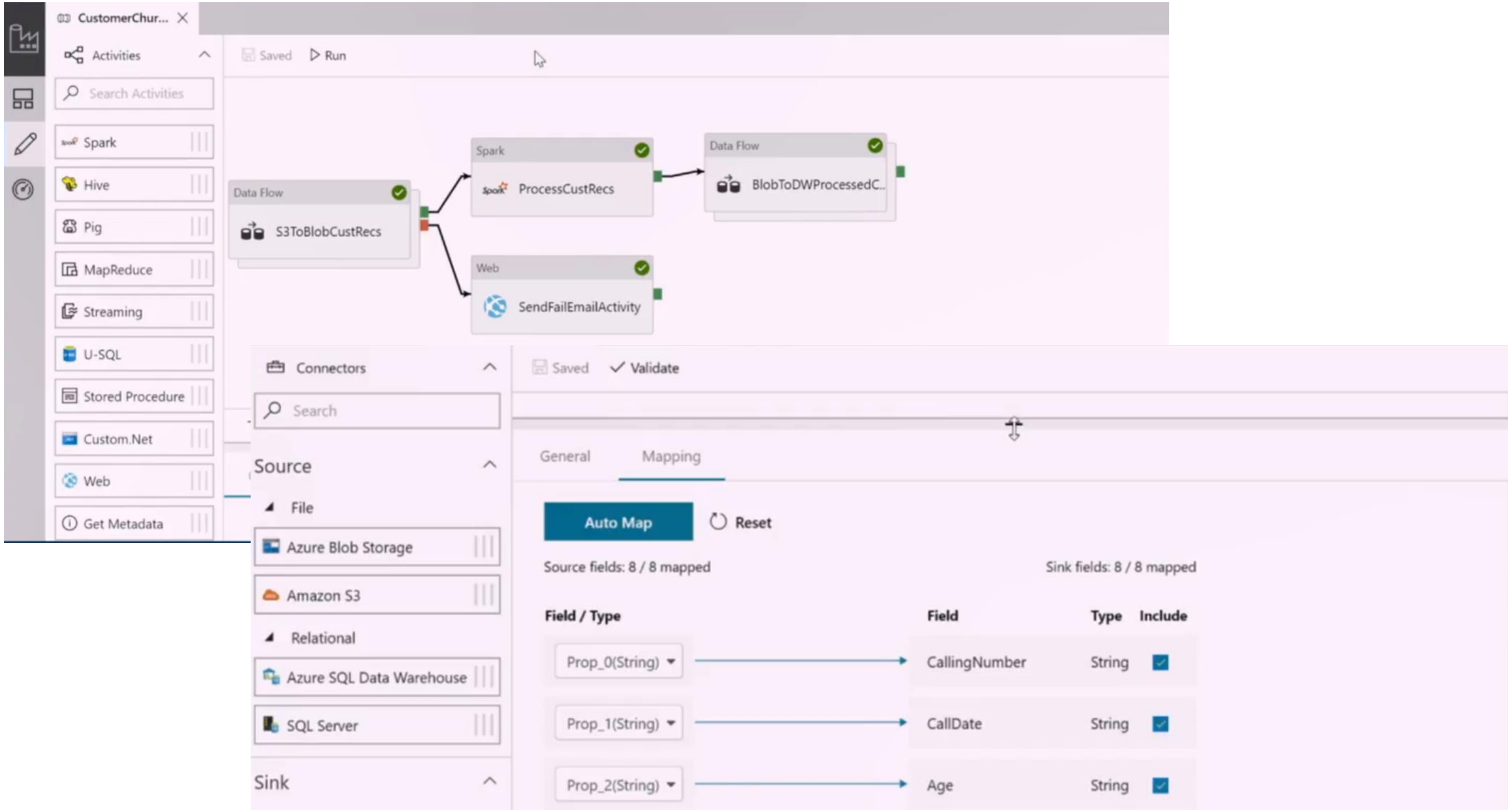
- Data movement activities : Copy Activity
- Data transformation activities :

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 V1 Pipelines



Azure Data Factory V2 Pipelines



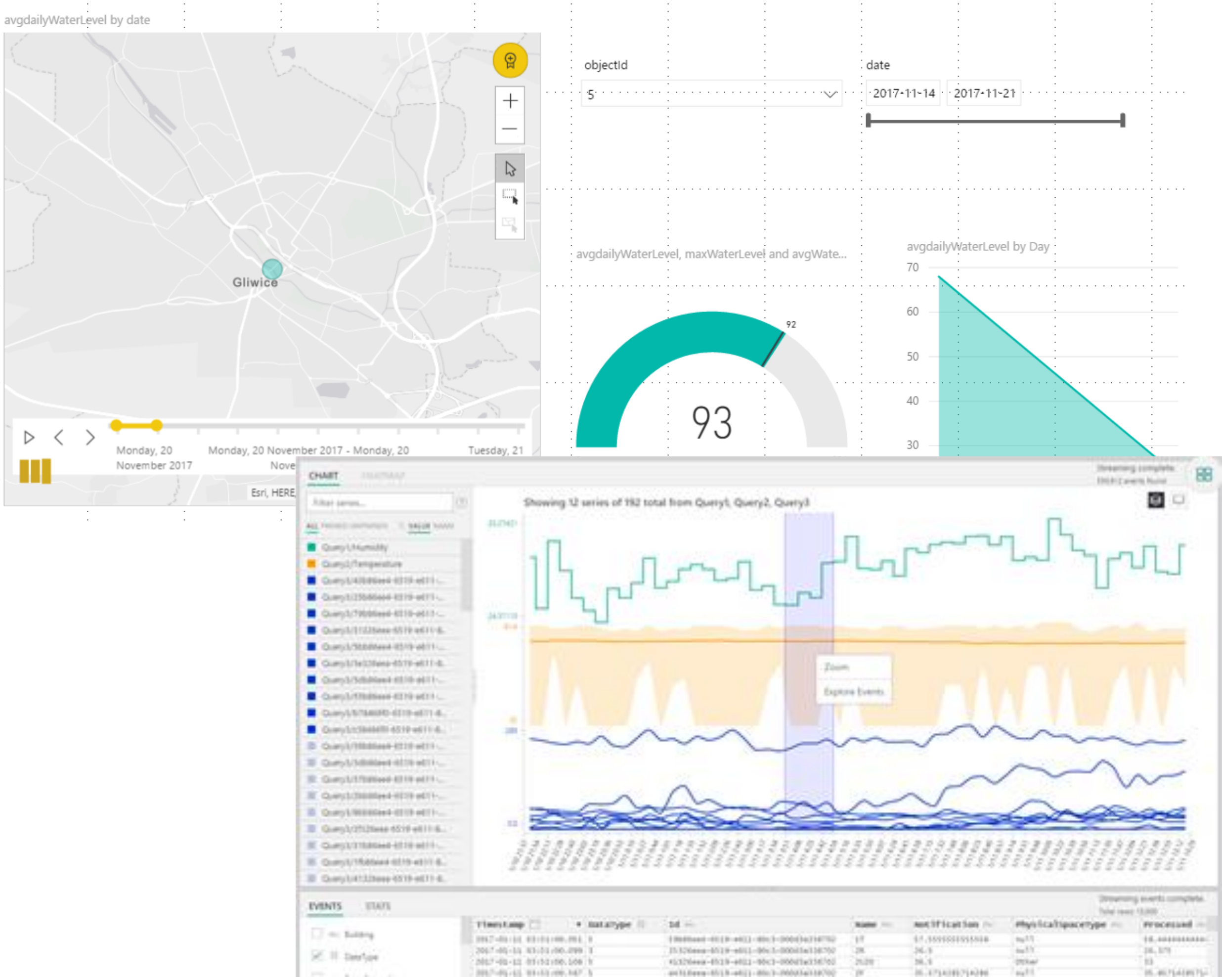
Source fields: 8 / 8 mapped

Sink fields: 8 / 8 mapped

Field / Type	Field	Type	Include
Prop_0(String)	CallingNumber	String	☑
Prop_1(String)	CallDate	String	☑
Prop_2(String)	Age	String	☑

Azure Serving Layer

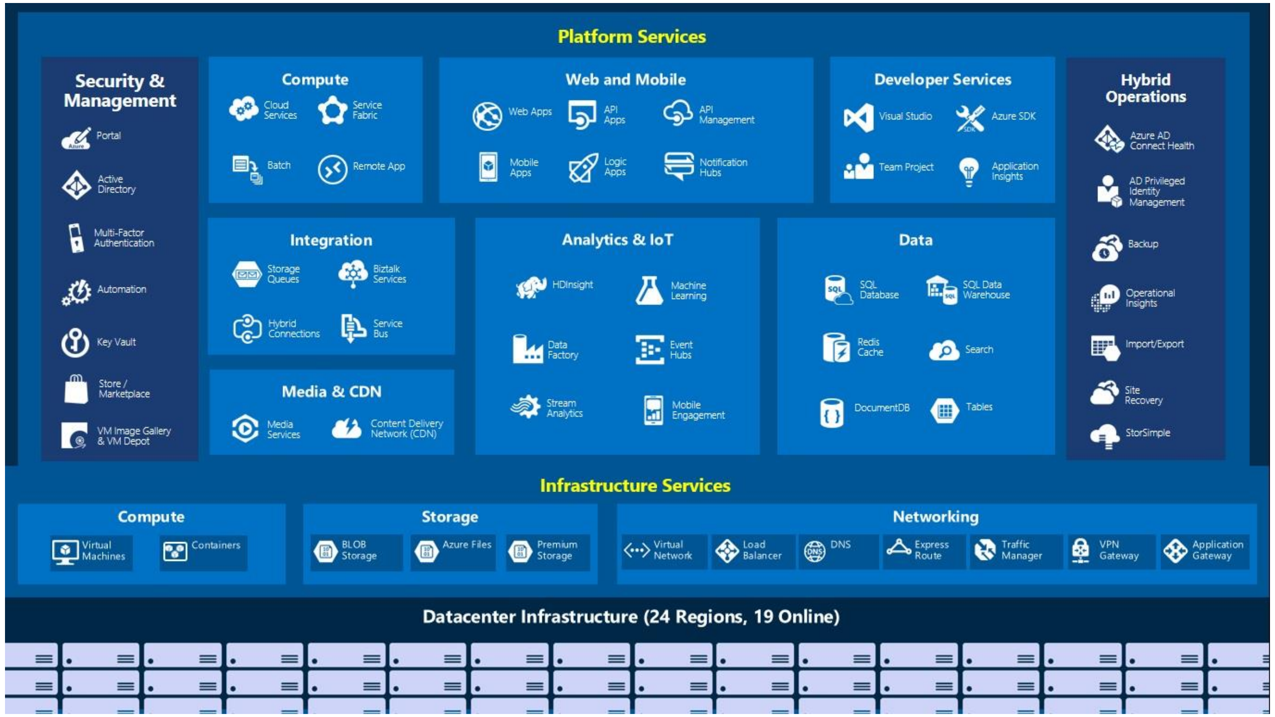
- › Power BI
- › Time Series Insight
- › Excel
- › API



Showing 12 series of 192 total from Query1, Query2, Query3

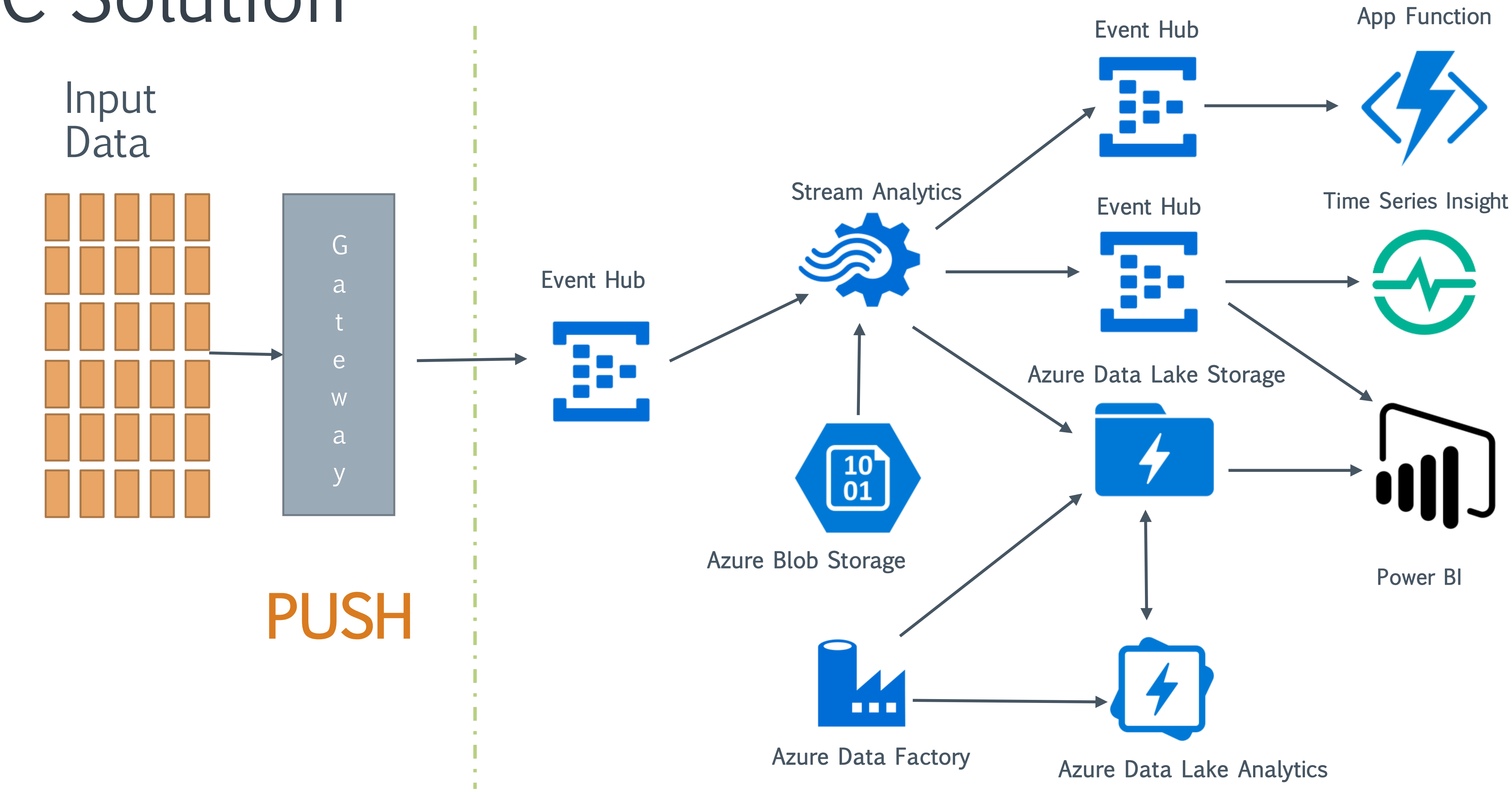
Azure as a Big Data Platform
What else?

- Azure Data Warehouse
- Azure Cosmos DB
- Azure Notebooks
- Azure Databricks
- IoT Hub (IoT Edge)
  - + Azure IoT Suite
- Azure Data Catalog
- Azure Cognitive Services
- Azure ML Studio
- Azure App Functions
- Azure AD
- Azure Batches



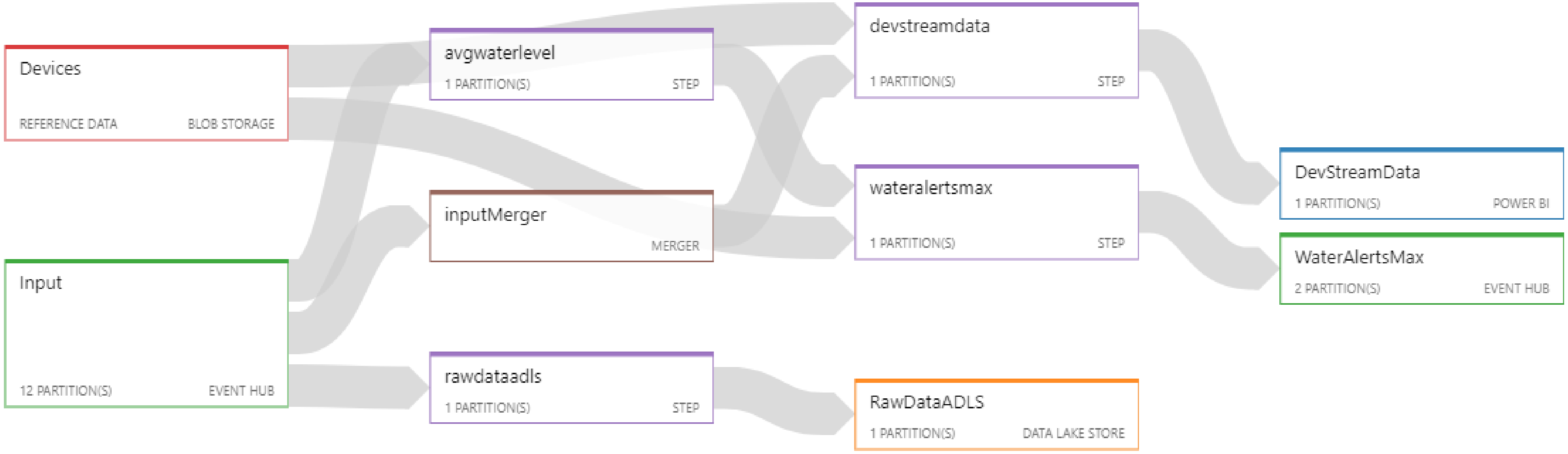
First Challenge Solution

PoC Solution



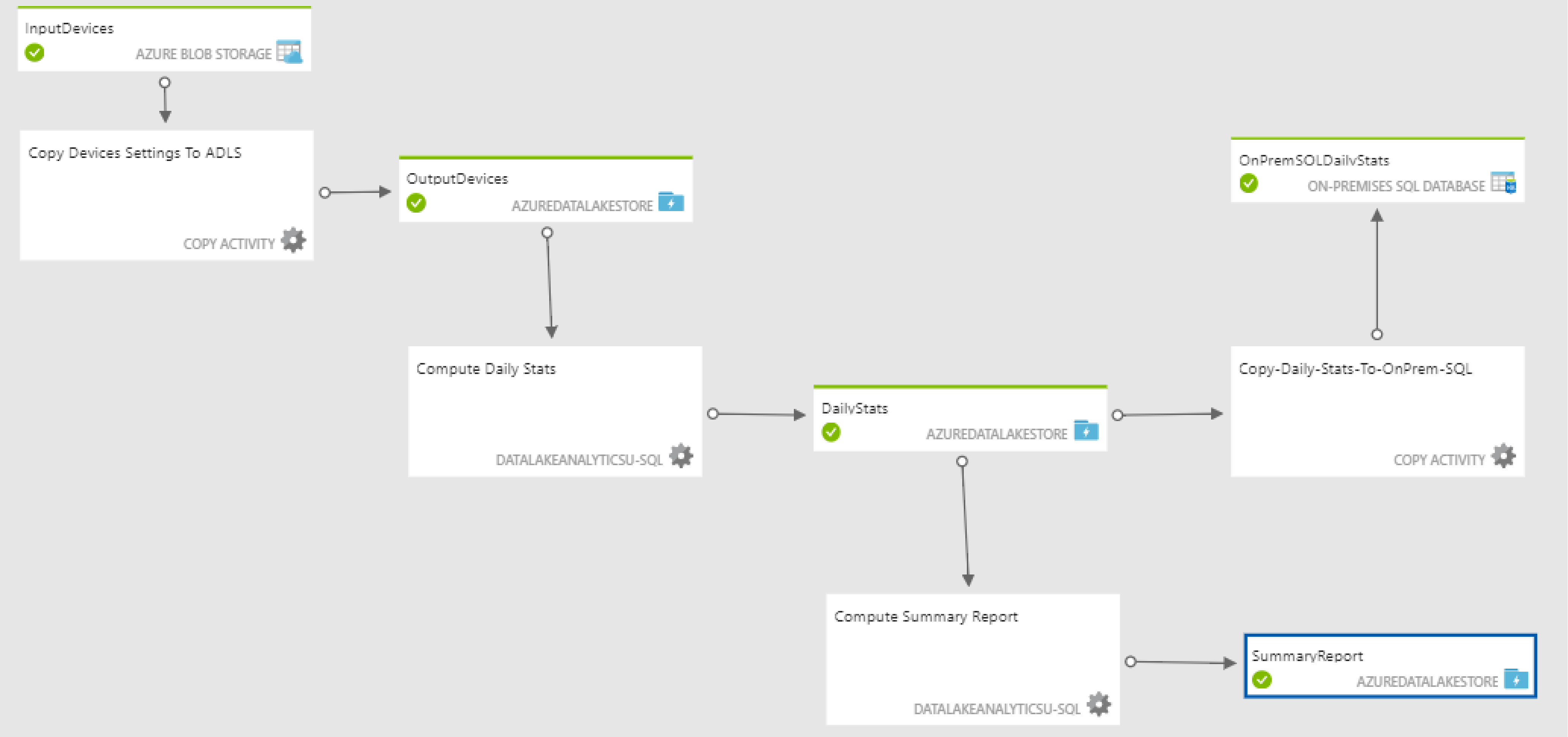
First Challenge Solution –Hot Path

› Azure Stream Analytics Job (Hot Path)



First Challenge Solution –Hot Path

- Azure Data Factory Pipeline (Cold Path)



DEMO AND Q&A

- Resources:
  https://docs.microsoft.com/en-us/azure/
  https://github.com/cloud4yourdata/demos/tree/develop
- Contact:
  tomasz.k.krawczyk@gmail.com
  tkrawczyk@future-processing.com