

Business Intelligence in Azure

SQL GRILLEN



Alex Whittles



Alex@PurpleFrogSystems.com

PurpleFrogSystems.com

PurpleFrogSystems.com/blog

[@PurpleFrogSys](https://twitter.com/PurpleFrogSys)

Alex Whittles



- SQL Relay Committee



SQLRelay.co.uk

- SQL Bits Committee



SQLBits.com

- Birmingham SQL UG



SQLMidlands.com

- Birmingham Azure UG



AzureBirmingham.uk

- MSc in Business Intelligence, CEng, CITP, FBCS, FIOEE, MIET, MIOD

- Run  BI Consultancy



Business Intelligence Consultancy

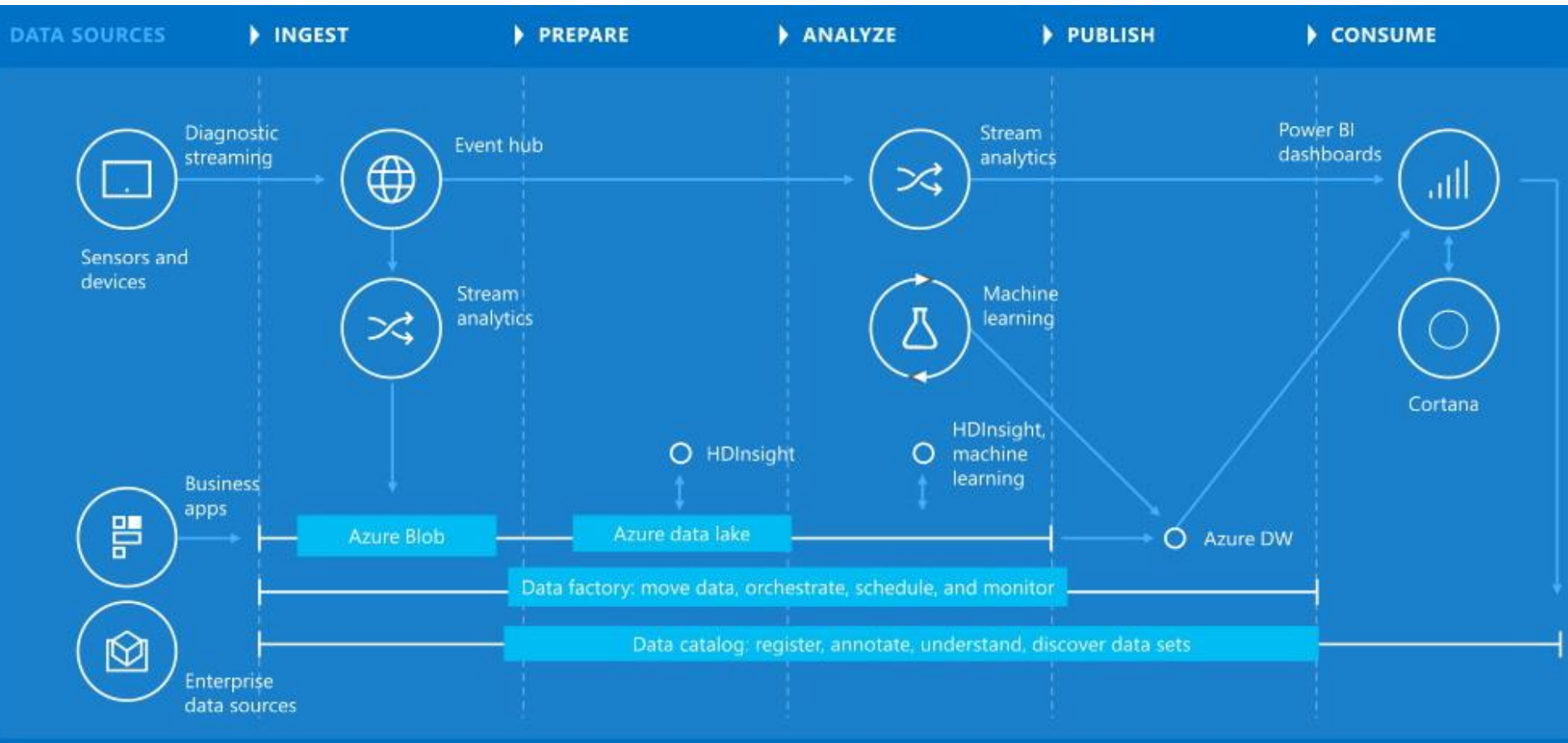
Data Modelling
Data Warehousing
OLAP Cubes

ETL Systems
Reporting Systems
Managed Service

Alex Whittles



Cortana Analytics Data Flow



Business Intelligence in Azure

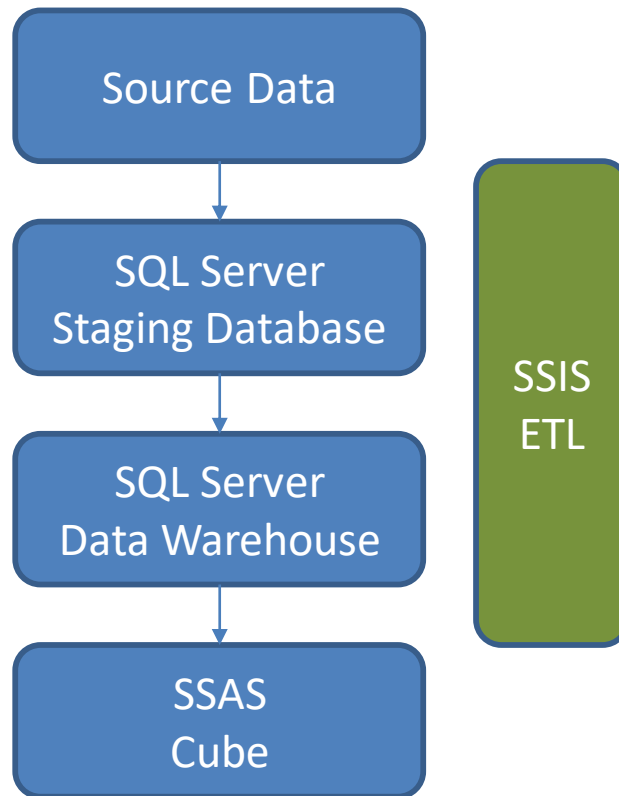
Options:

1. Migrating on-premise BI solutions to Azure
2. Batch load BI systems in Azure
3. Real-time BI systems in Azure

For each:

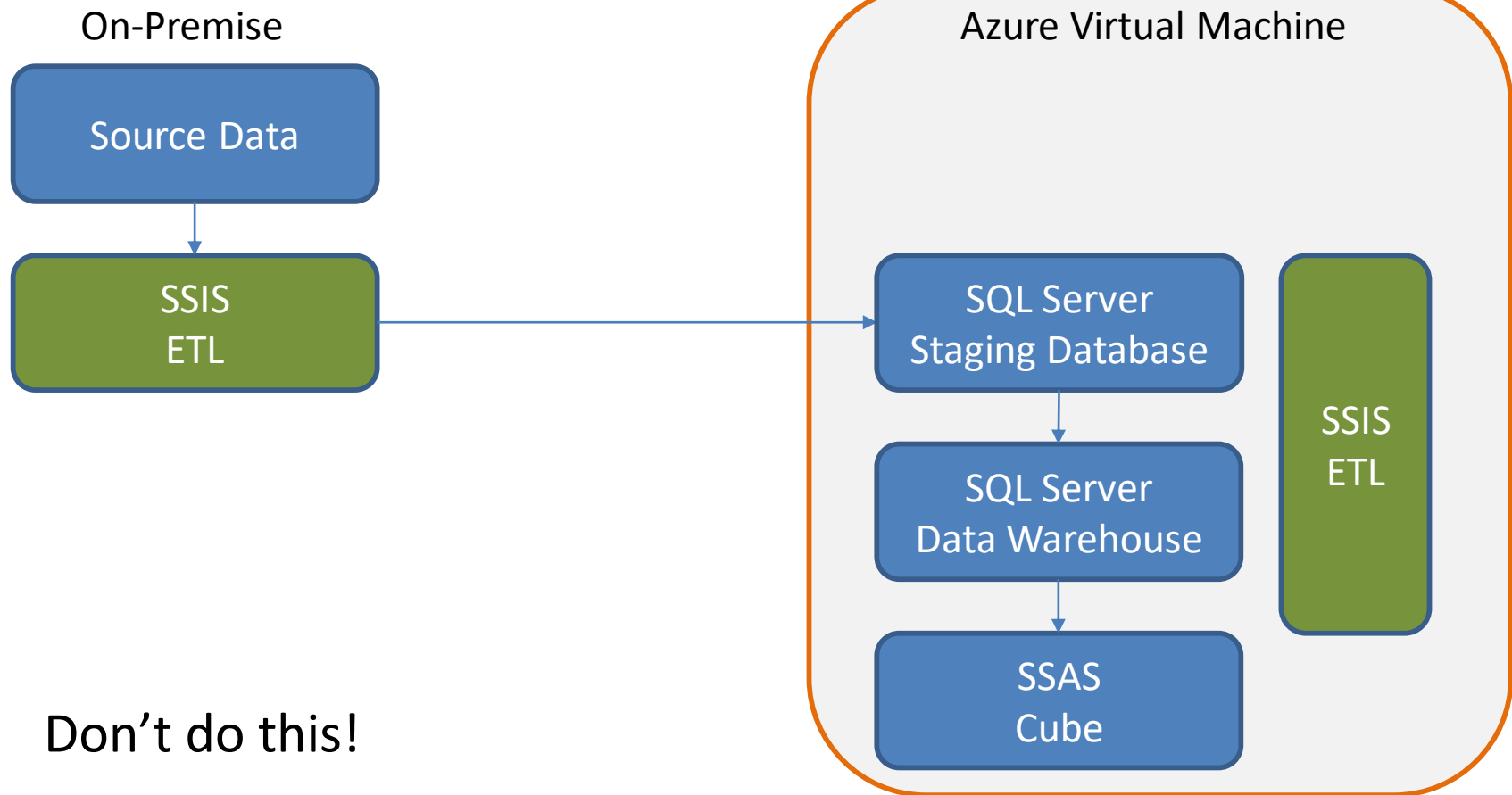
Staging / ETL / Data Warehouse / Cube

On-Premise Batch Load BI



Azure IAAS Solution

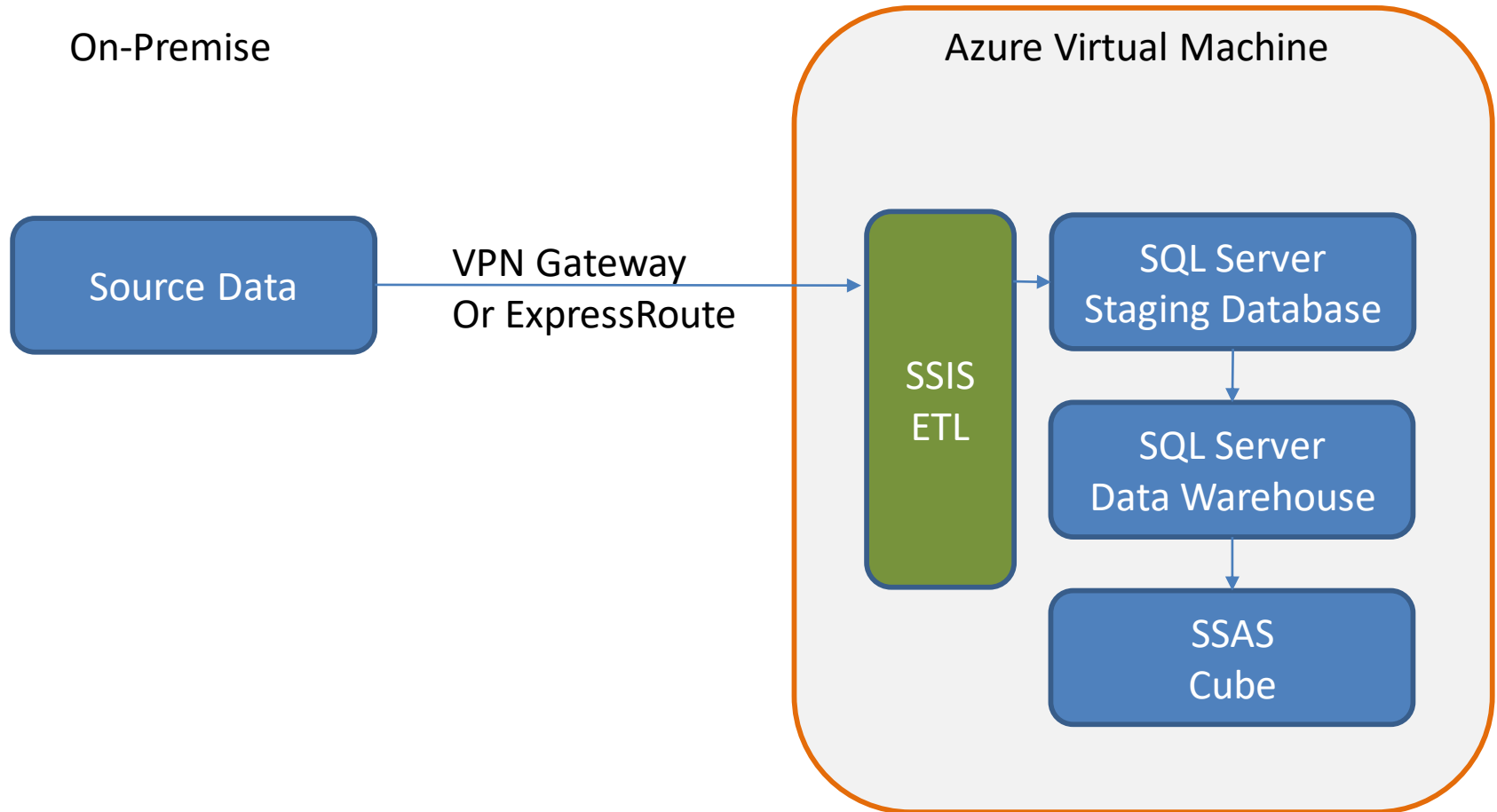
[Infrastructure As A Service]



Better Azure IAAS Solution

[Infrastructure As A Service]

On-Premise

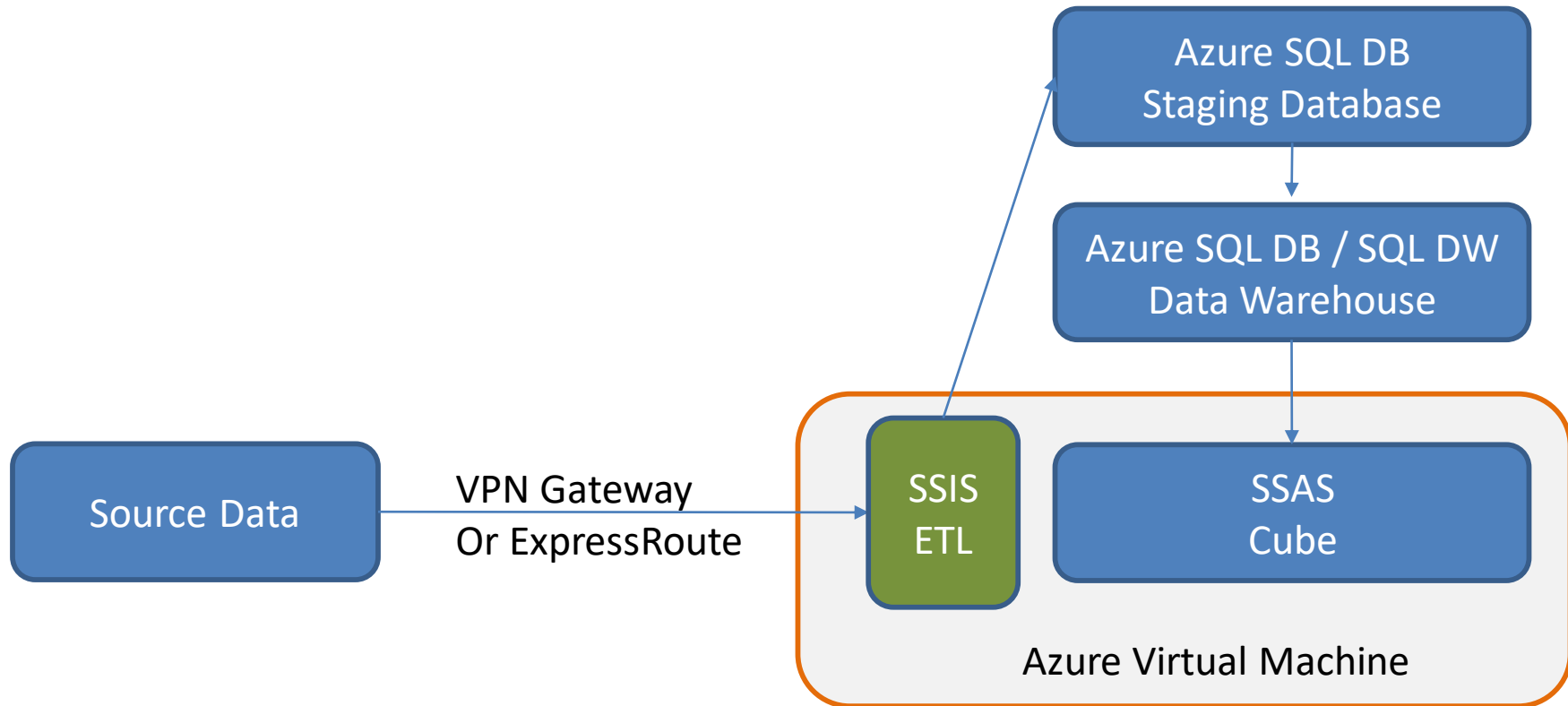


PAAS / IAAS Hybrid

[Platform As A Service]

On-Premise

Azure

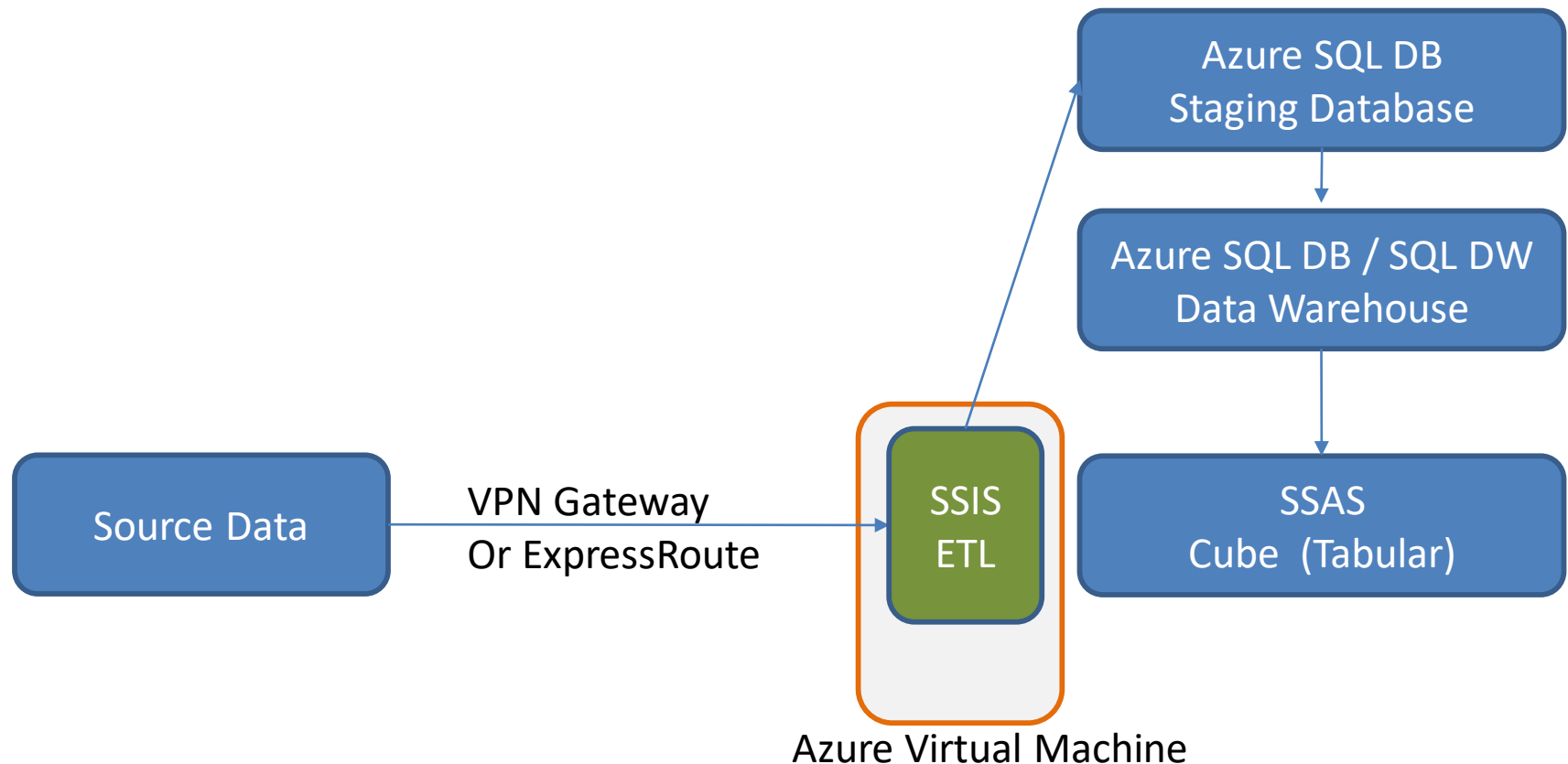


PAAS / IAAS Hybrid

[Platform As A Service]

On-Premise

Azure



PAAS SSAS Tabular

Data + Analytics

intelligent bot development powered by the Microsoft Bot

Data Lake Analytics
Big data analytics made easy

Time Series Insights (preview)
Azure Time Series Insights is a fully managed analytics, storage, and visualization service that makes it

Data Lake Store
Hyper-scale repository for big data analytic workloads

Data Factory
Transform data into trusted information

Data Catalog
Data source discovery to get more value from existing enterprise data assets

Power BI Embedded
Embed fully interactive, stunning data visualizations in your applications

Log Analytics
Collect, search and visualize machine data from on-premises and cloud

Analysis Services
Proven analytical engine in the cloud.

Analysis Services

Server name

Subscription

Resource group

Location

Pricing tier

Administrator

Storage Settings

Storage key expiration

☐ Pin to dashboard

Create

Choose your pricing tier

Browse the available plans

B1 Basic	B2 Basic	S0 Standard
40 Query Processing Units	80 Query Processing Units	40 Query Processing Units
Up to 10 GB Cache	Up to 20 GB Cache	Up to 10 GB Cache
Dedicated service	Dedicated service	Dedicated service
SSL	SSL	SSL
119.22 GBP/MONTH (ESTIMATED)	238.44 GBP/MONTH (ESTIMATED)	335.48 GBP/MONTH (ESTIMATED)

S1 Standard	S2 Standard	S4 Standard
100 Query Processing Units	200 Query Processing Units	400 Query Processing Units
Up to 25 GB Cache	Up to 50 GB Cache	Up to 100 GB Cache
Dedicated service	Dedicated service	Dedicated service
SSL	SSL	SSL
562.83 GBP/MONTH (ESTIMATED)	1,125.66 GBP/MONTH (ESTIMATED)	2,248.54 GBP/MONTH (ESTIMATED)

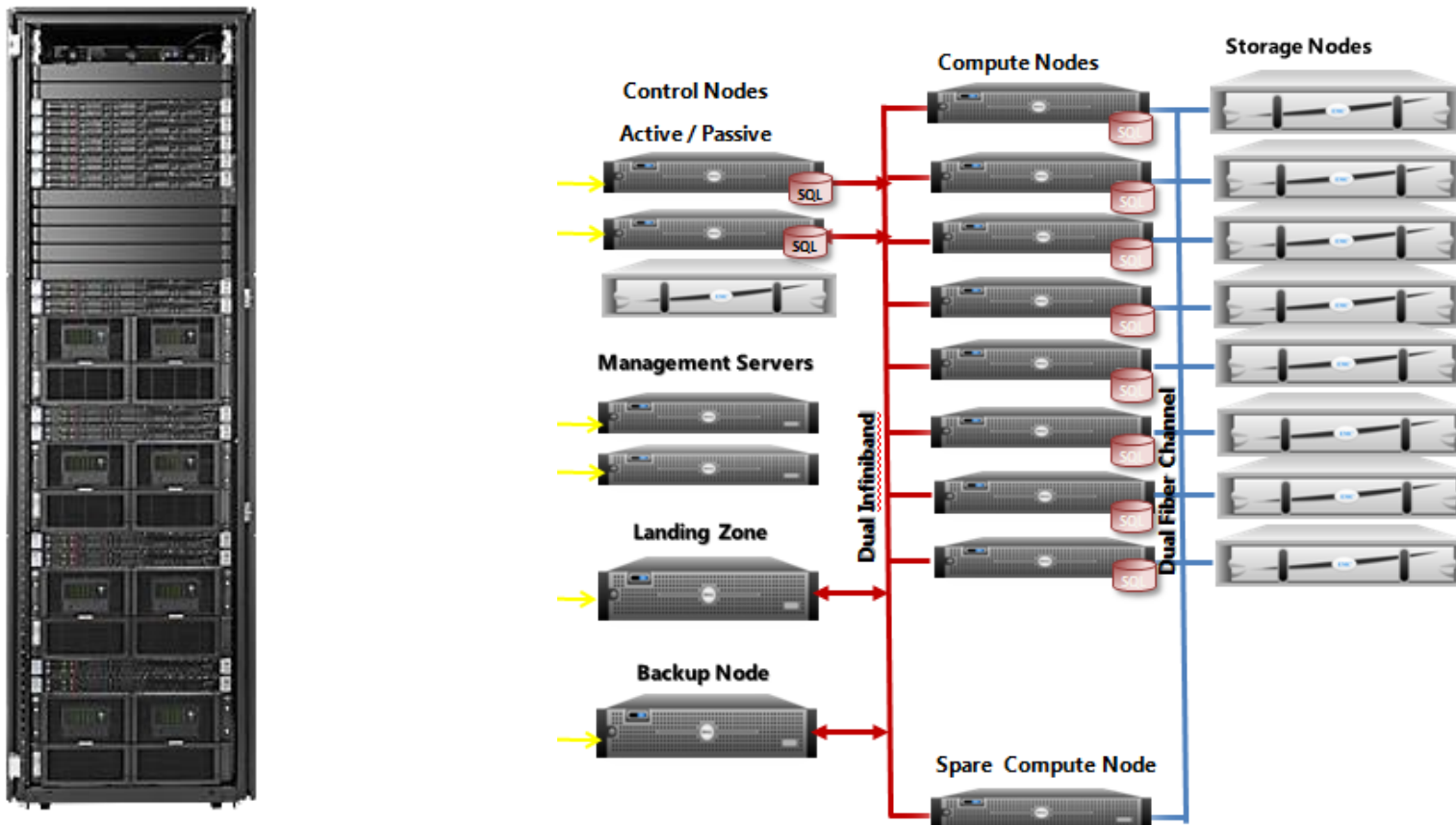
D1 Developer
20 Query Processing Units
Up to 3 GB Cache
Shared infrastructure
SSL
36.60 GBP/MONTH (ESTIMATED)

Select

PDW / APS / SQL DW

PDW / APS = Parallel Data Warehouse / Analytics Platform System
appliance, on premise

SQL DW = Same concept but in Azure. Each node is an Azure SQL DB



Azure SQL Data Warehouse

Microsoft Azure > New > Databases > SQL Data Warehouse

Search the marketplace

MARKETPLACE See all

- Compute
- Networking
- Storage
- Web + mobile
- Databases
- Intelligence + analytics
- Internet of things
- Enterprise integration
- Security + identity
- Developer tools
- Monitoring + management
- Add-ons
- Containers

RECENT

- SQL Database Microsoft
- Data Lake Analytics (preview) Microsoft

FEATURED APPS See all

- SQL Database**
Scalable and managed relational database service for modern business-class apps.
- SQL Data Warehouse**
Fully elastic, managed, and parallelized relational database. Analyze and scale in seconds.
- SQL Elastic Database Pool**
Scalable and managed relational elastic database pool.
- SQL Server 2016 RTM Enterprise on Windows Server 2012 R2**
Enterprise version of SQL Server
- DocumentDB (NoSQL)**
Scalable and managed NoSQL document database service for modern cloud applications.
- DocumentDB (NoSQL) - Protocol Support for MongoDB (preview)**
Fully managed NoSQL document
- Redis Cache**
Distributed, in-memory Redis Cache service for modern cloud

SQL Data Warehouse

SQL Data Warehouse has moved to General Availability pricing on September 1st, 2016.

* Database name
SQLRelay

* Subscription
Visual Studio Enterprise with MSDN

* Resource group
Create new Use existing
Default-SQL-NorthEurope

* Select source
Blank database

* Server
wvh4h34u93 (North Europe)

* Collation
SQL_Latin1_General_CP1_CI_AS

Performance
400
400 DWU X 0.0074 = **2.96** GBP / HOUR
[Learn more about pricing.](#)

Azure SQL DW – Table Distribution

Round Robin

```
CREATE TABLE [dbo].[FactInternetSales]
(
    [ProductKey]          int          NOT NULL
,   [OrderDateKey]       int          NOT NULL
,   [CustomerKey]        int          NOT NULL
,   [PromotionKey]       int          NOT NULL
,   [SalesOrderNumber]   nvarchar(20) NOT NULL
,   [OrderQuantity]      smallint     NOT NULL
,   [UnitPrice]          money        NOT NULL
,   [SalesAmount]        money        NOT NULL
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX
,   DISTRIBUTION = ROUND_ROBIN
)
;
```

HASH

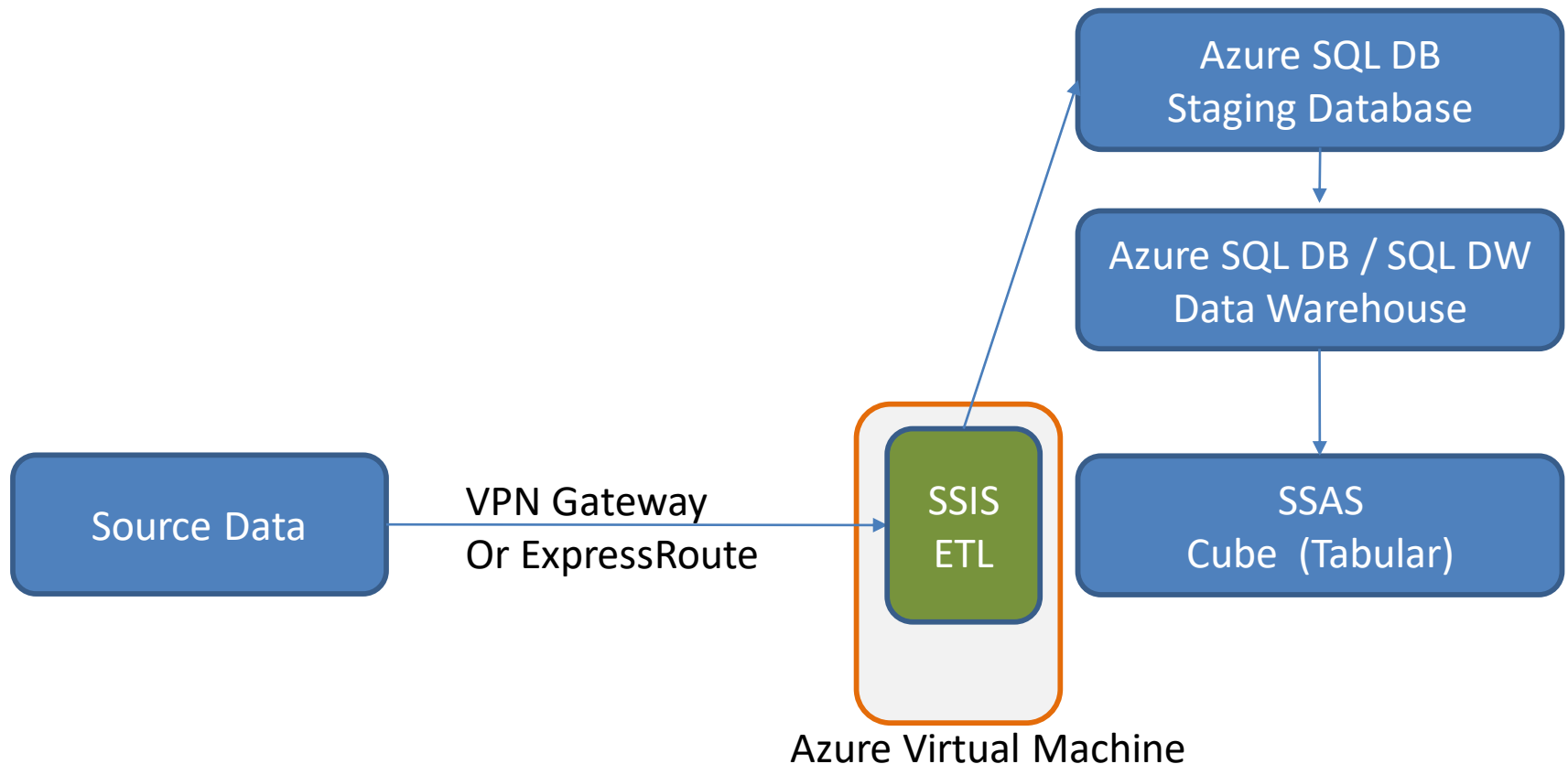
```
CREATE TABLE [dbo].[FactInternetSales]
(
    [ProductKey]          int          NOT NULL
,   [OrderDateKey]       int          NOT NULL
,   [CustomerKey]        int          NOT NULL
,   [PromotionKey]       int          NOT NULL
,   [SalesOrderNumber]   nvarchar(20) NOT NULL
,   [OrderQuantity]      smallint     NOT NULL
,   [UnitPrice]          money        NOT NULL
,   [SalesAmount]        money        NOT NULL
)
WITH
(
    CLUSTERED COLUMNSTORE INDEX
,   DISTRIBUTION = HASH([ProductKey])
)
;
```

PAAS / IAAS Hybrid

[Platform As A Service]

On-Premise

Azure

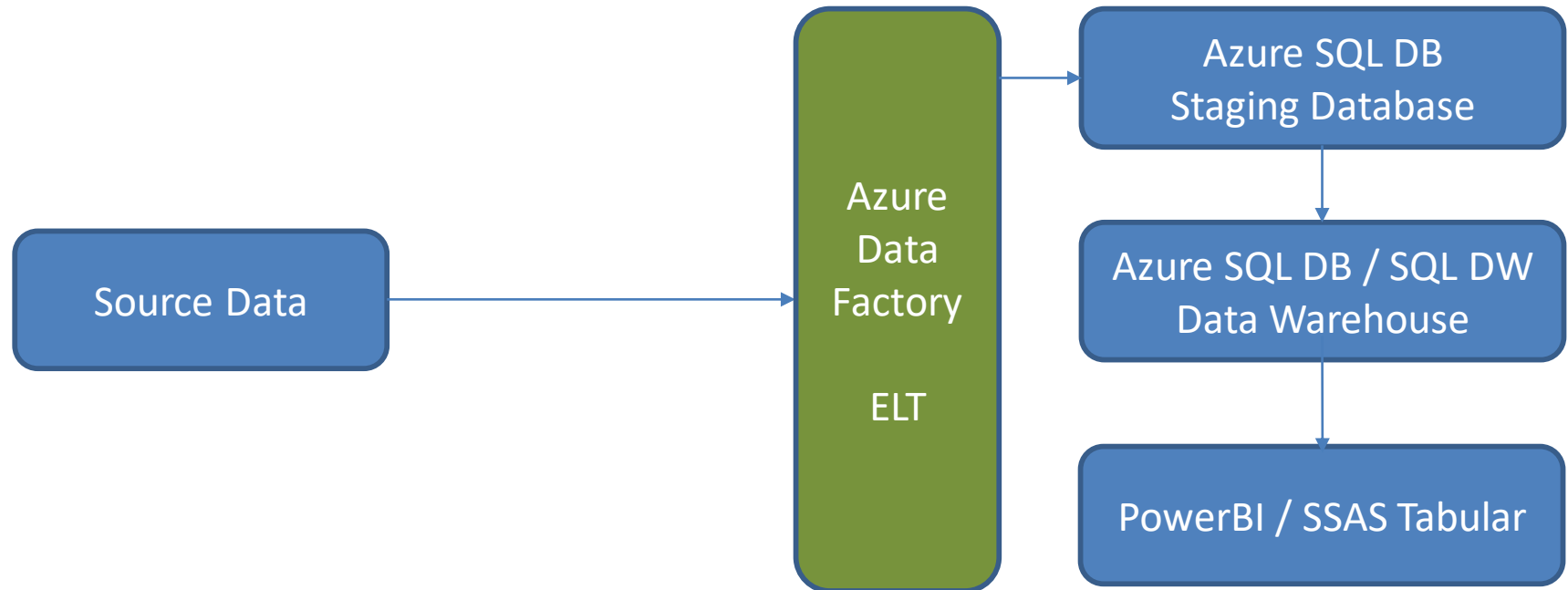


PAAS Solution

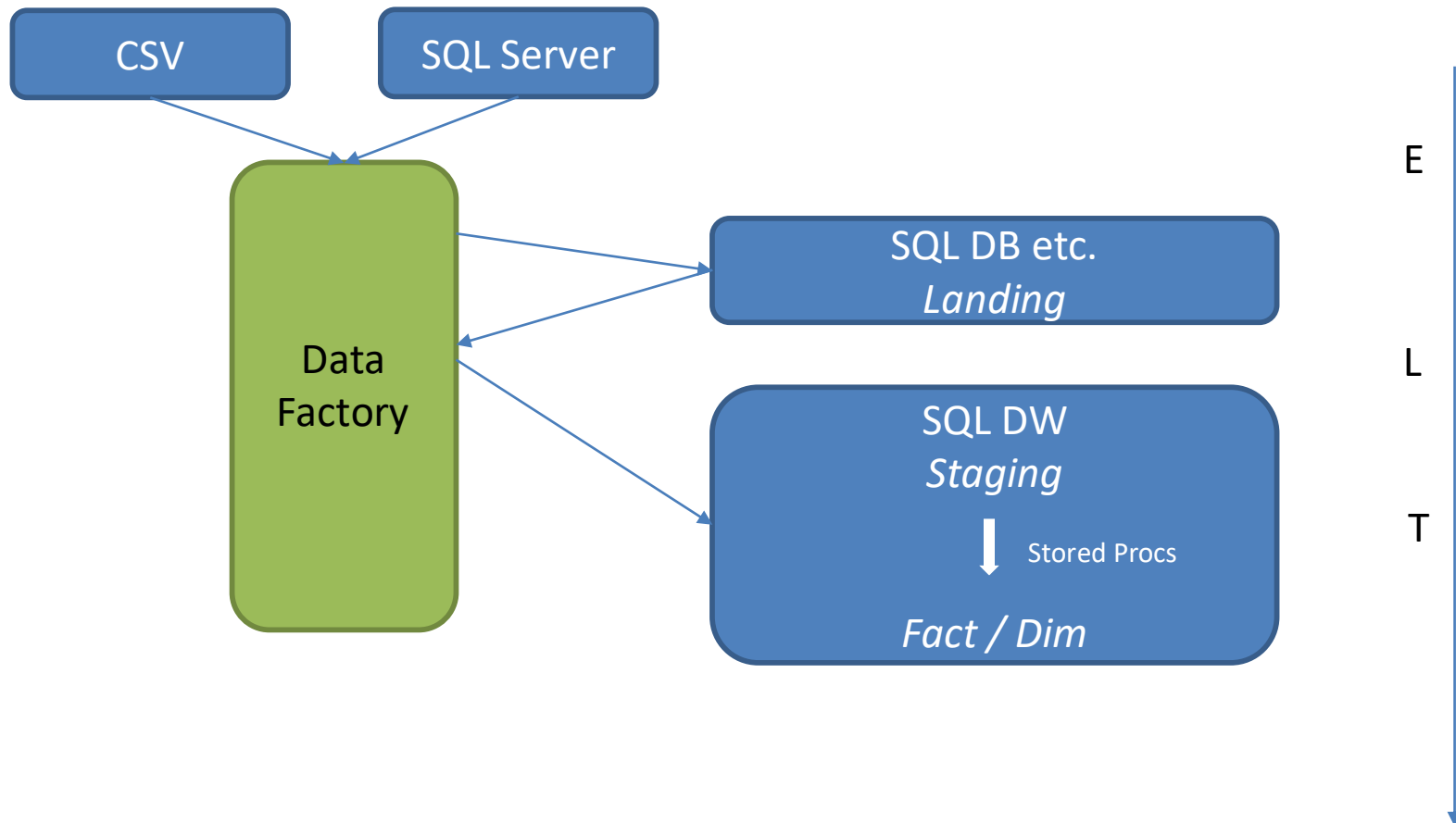
[Platform As A Service]

On-Premise

Azure



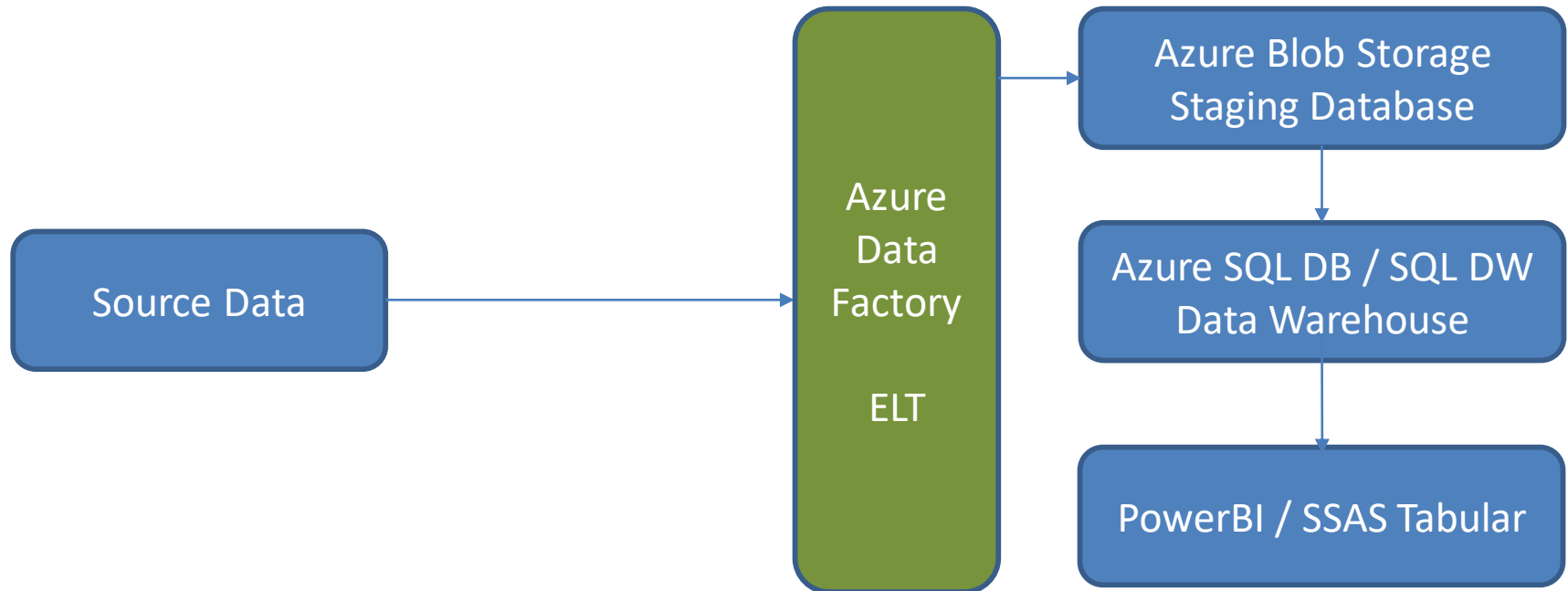
Example ELT Project



Cheaper Staging – Blob/Table

On-Premise

Azure



Blob / Table Storage

Blob

- Azure File System
- Any file, any format
- Unstructured
- Cheap

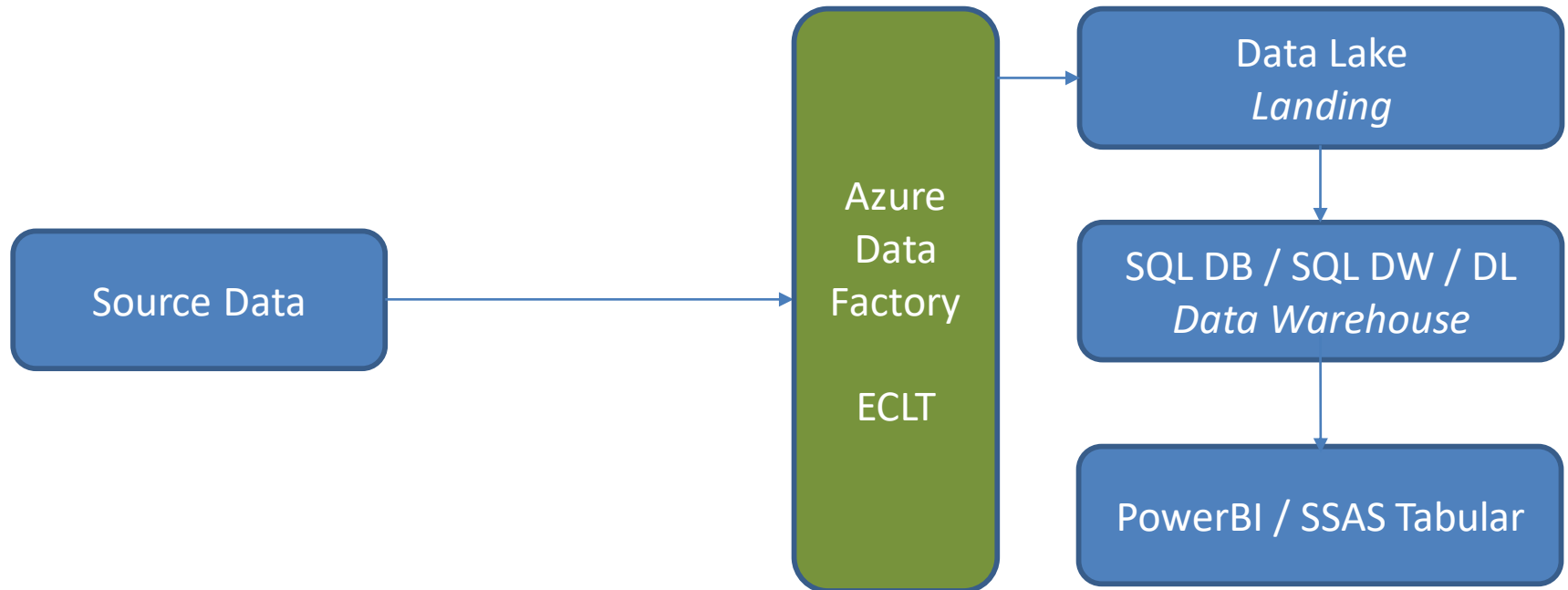
Table Storage

- Simple tables
- No real indexing
- No foreign keys
- Basic data types
- Basic querying
- Structured
- Cheap

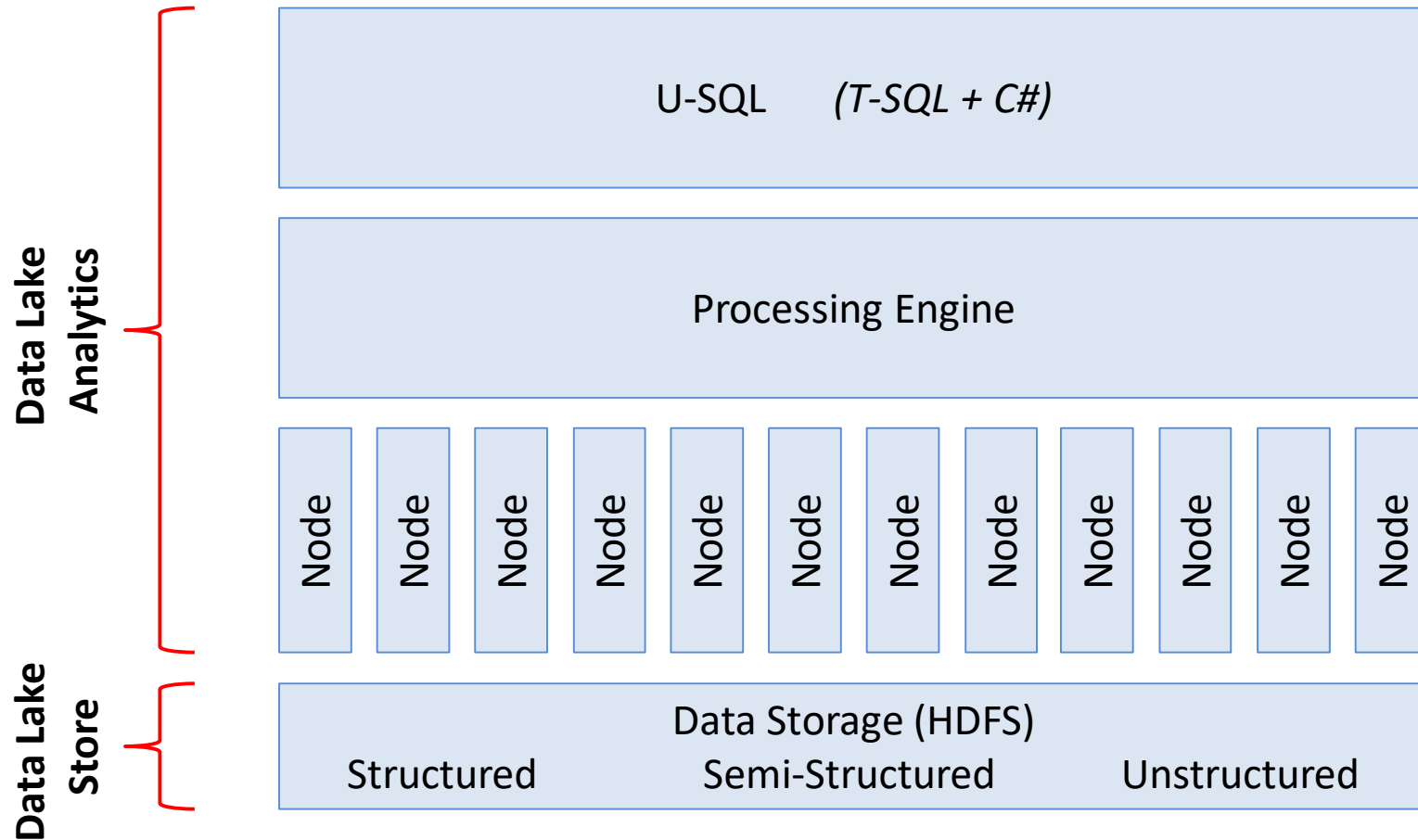
Integrate Data Lake

On-Premise

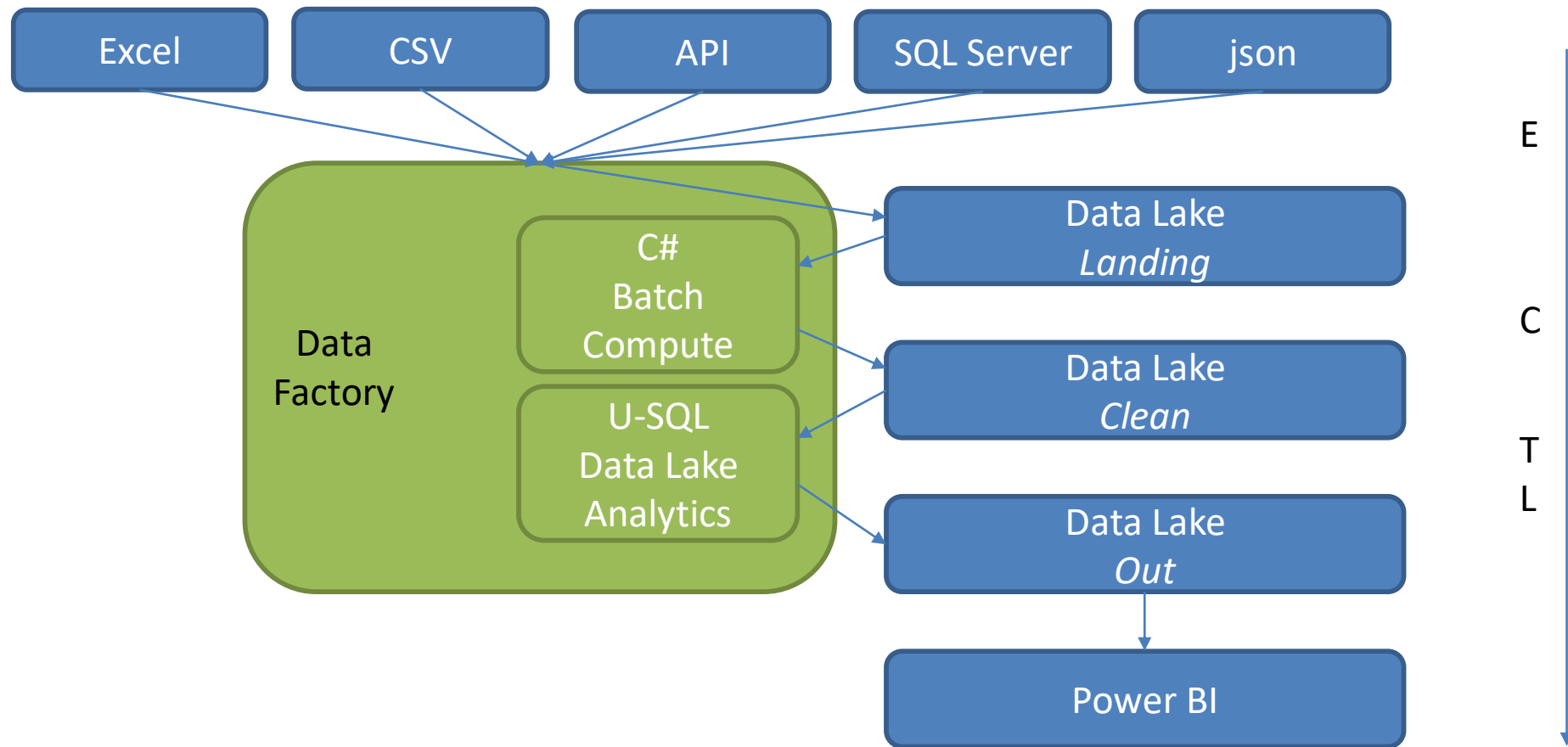
Azure



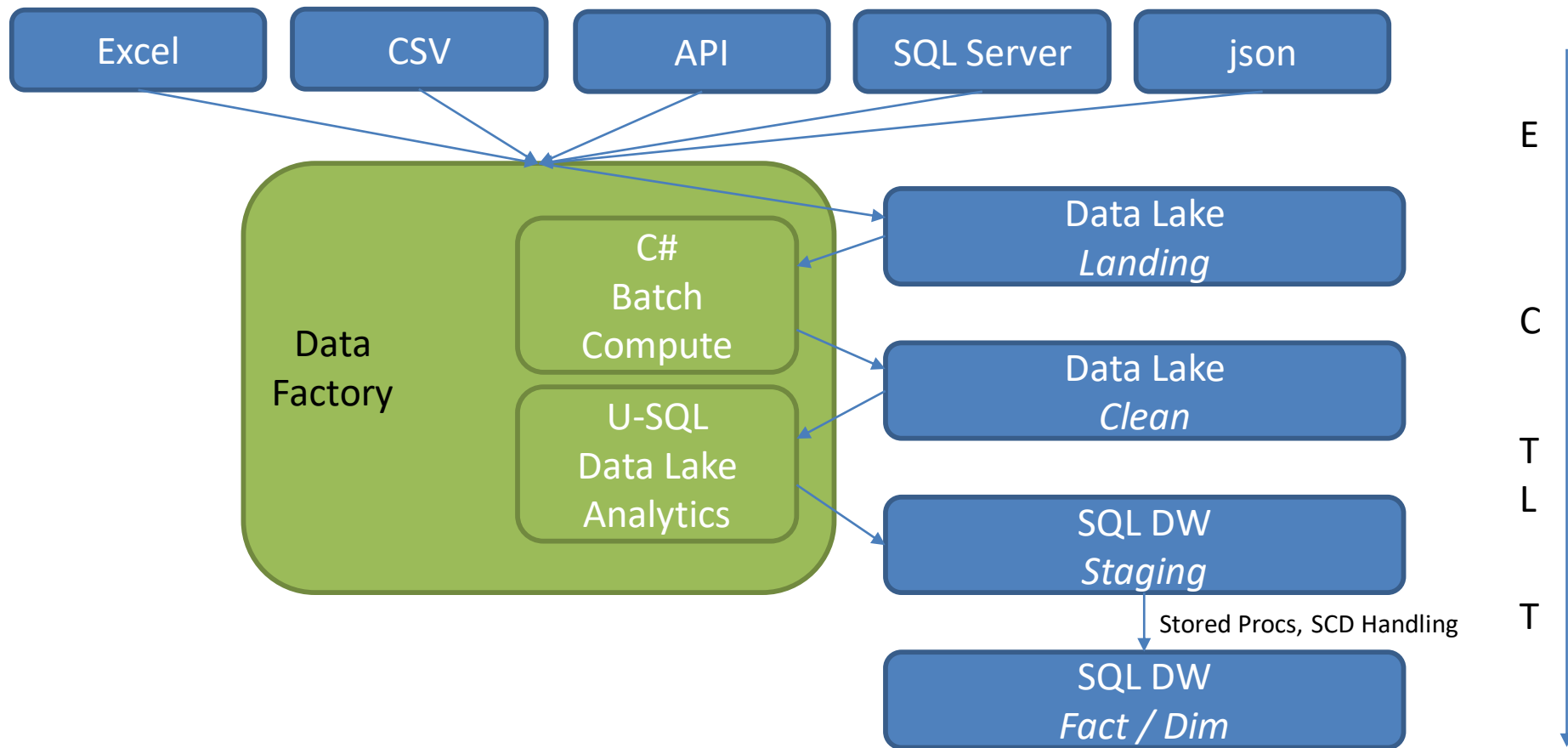
Data Lake



Example ECTL Project



Example ECTLT Project



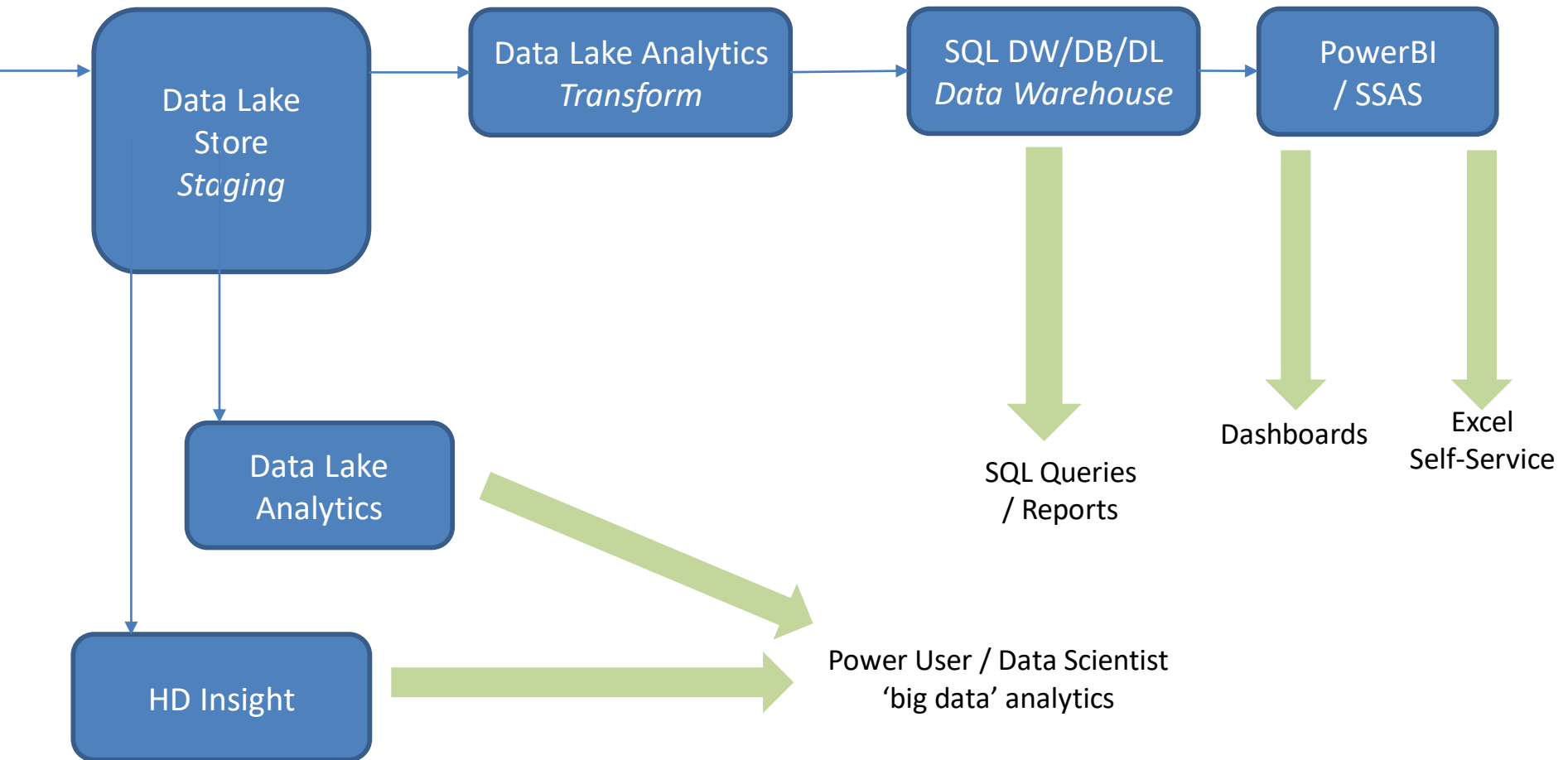
ETLT? Or ECLT? Or ECTLT?!

- Extract – get the data
- Clean – prepare the data files
- Transform – create facts/dims?
- Load – load into DW/etc.
- Transform – create facts/dims?
– SCD handling

Choose the right combination for your project

Data Lake

Bridge between Normal & Big Data



Data Lake Analytics Vs HDInsight

Data Lake

- U-SQL
- (T-SQL + C#)

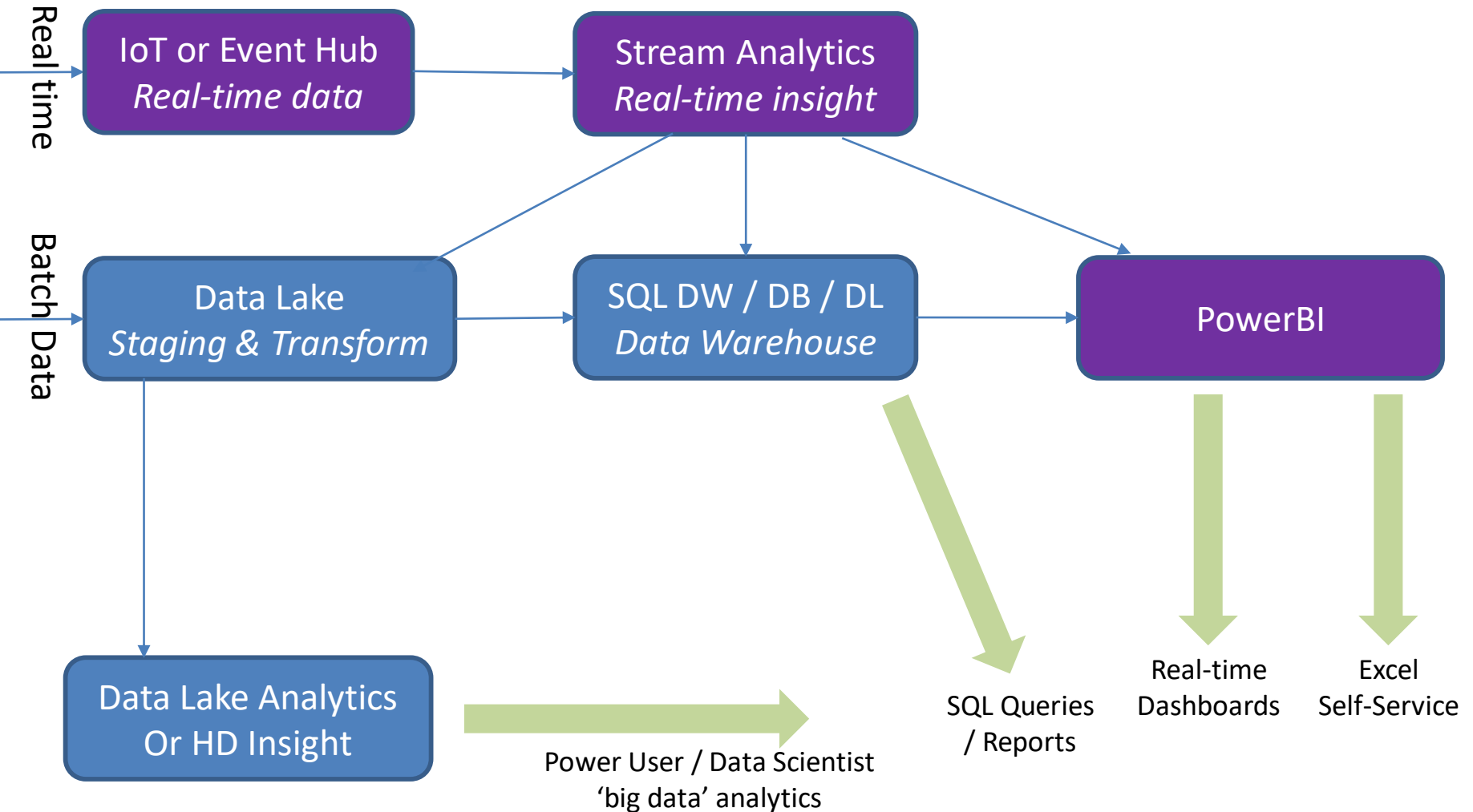
```
@CustomersFullAddress =  
    SELECT CustomerName,  
           Address.Split(',')[0] AS Address1,  
           Address.Split(',')[1] AS Address2,  
           Address.Split(',')[2] AS Address3,  
           myUSQLTest.Udfs.ValidatePostCode(PostCode) AS PostCode,  
           ((Gender=="M" || Gender=="F") ? Gender : "-") AS Gender,  
           myUSQLTest.Udfs.ValidateDate(DoB) AS DoB,  
           ROW_NUMBER() OVER (PARTITION BY CustomerName ORDER BY LastUpdated) AS RN  
    FROM @Customers;
```

HDInsight

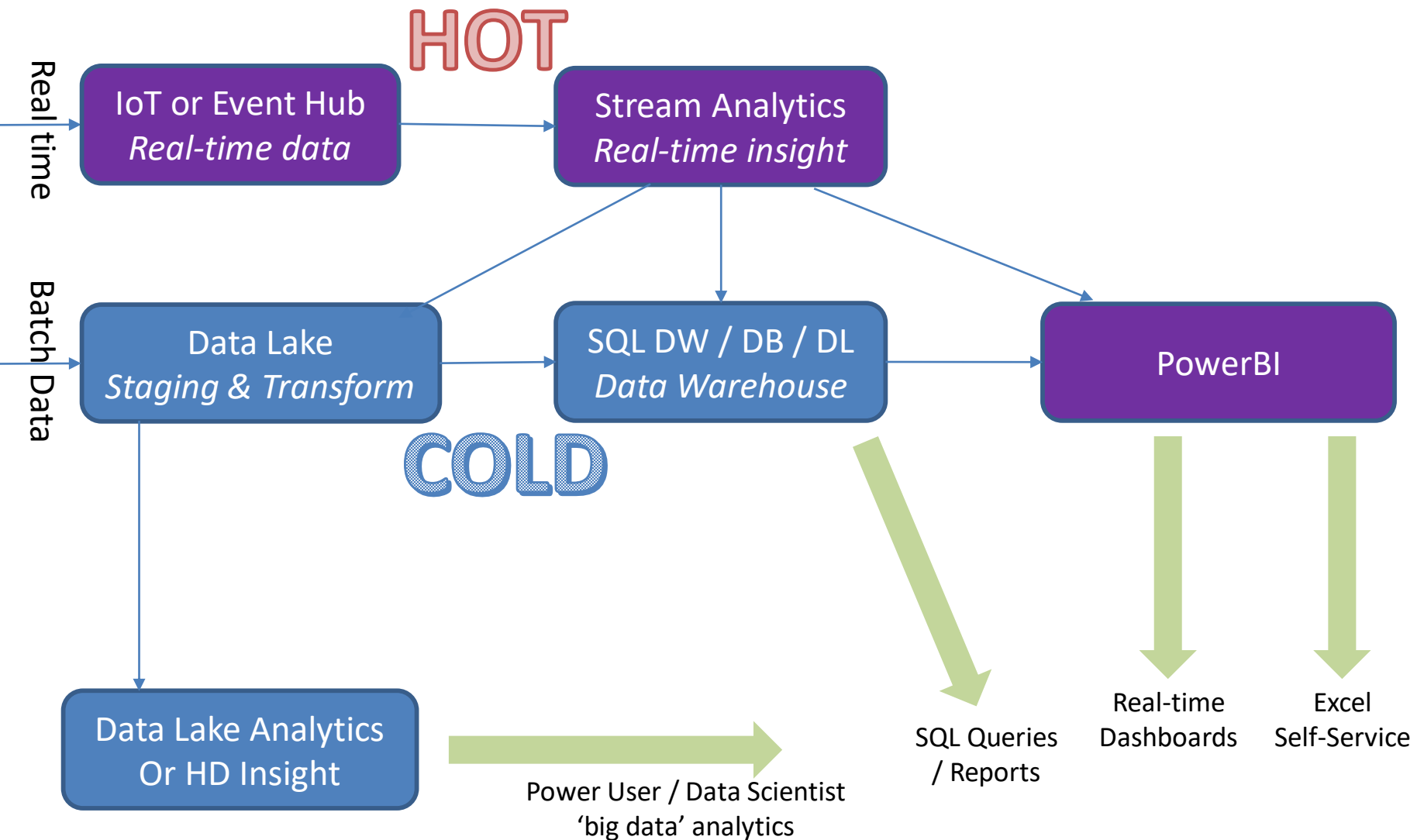
- Hadoop
- Map-Reduce
- Hive/Pig?

```
public static class IntSumReducer  
    extends Reducer<Text,IntWritable,Text,IntWritable> {  
    private IntWritable result = new IntWritable();  
  
    public void reduce(Text key, Iterable<IntWritable> values,  
                      Context context  
                      ) throws IOException, InterruptedException {  
        int sum = 0;  
        for (IntWritable val : values) {  
            sum += val.get();  
        }  
        result.set(sum);  
        context.write(key, result);  
    }  
}
```

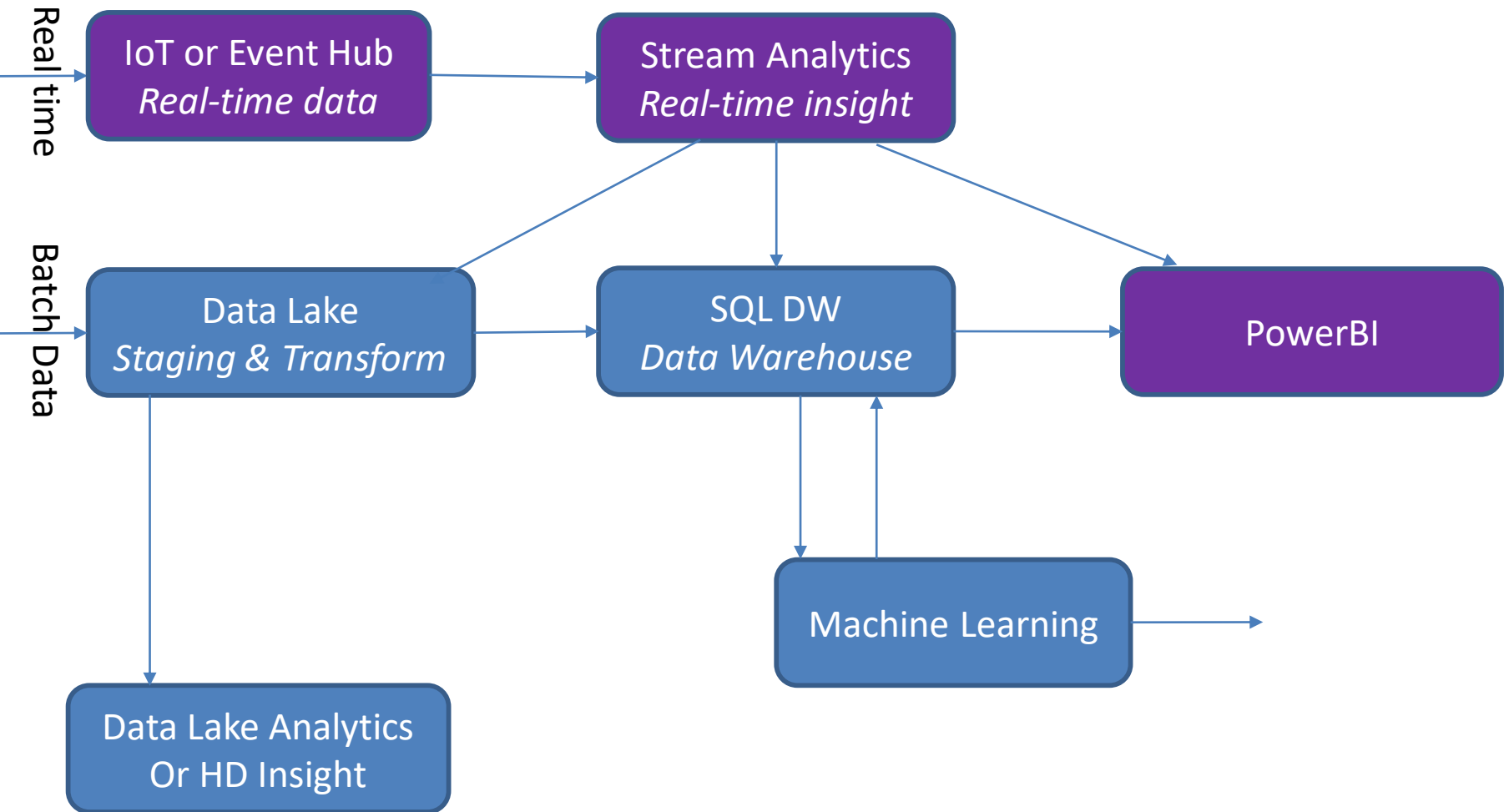
Real Time BI



Lambda Architecture



Machine Learning



Azure Machine Learning

The screenshot displays the Azure Machine Learning Studio interface. On the left is a sidebar with a search bar and a list of experiment items: Saved Datasets, Trained Models, Data Format Conversions, Data Input and Output, Data Transformation, Feature Selection, Machine Learning, OpenCV Library Modules, Python Language Modules, R Language Modules, Statistical Functions, Text Analytics, Deprecated, and Web Service. The main workspace is titled "Sample 5: Train, Test, Evaluate for Binary Classification" and is in "draft" status. It shows a workflow diagram with the following steps: "Adult Census Income Binary ..." (input), "Clean Missing Data" (checked), "Project Columns" (checked), "Split" (checked), "Two-Class Boosted Decision T..." (checked), "Train Model" (checked), "Permutation Feature Importa..." (checked), "Score Model" (checked), and "Evaluate Model" (checked). Arrows indicate the flow from the input dataset through the cleaning and transformation steps, then splitting into training and testing sets. The training set is used to train a model, which is then evaluated. The testing set is used to score the trained model. At the bottom, there is a toolbar with icons for "NEW", "VIEW RUN HISTORY", "SAVE", "SAVE AS", "DISCARD CHANGES", "REFRESH", "CANCEL", "RUN", "PREPARE WEB SERVICE", "PUBLISH TO GALLERY", and "UPDATE SCORING EXPERIMENT".

Search experiment items

Training experiment | Scoring experiment

Sample 5: Train, Test, Evaluate for Binary Classification In draft

Properties

Adult Census Income Binary ...

Clean Missing Data ✓

Project Columns ✓

Split ✓

Two-Class Boosted Decision T... ✓

Train Model ✓

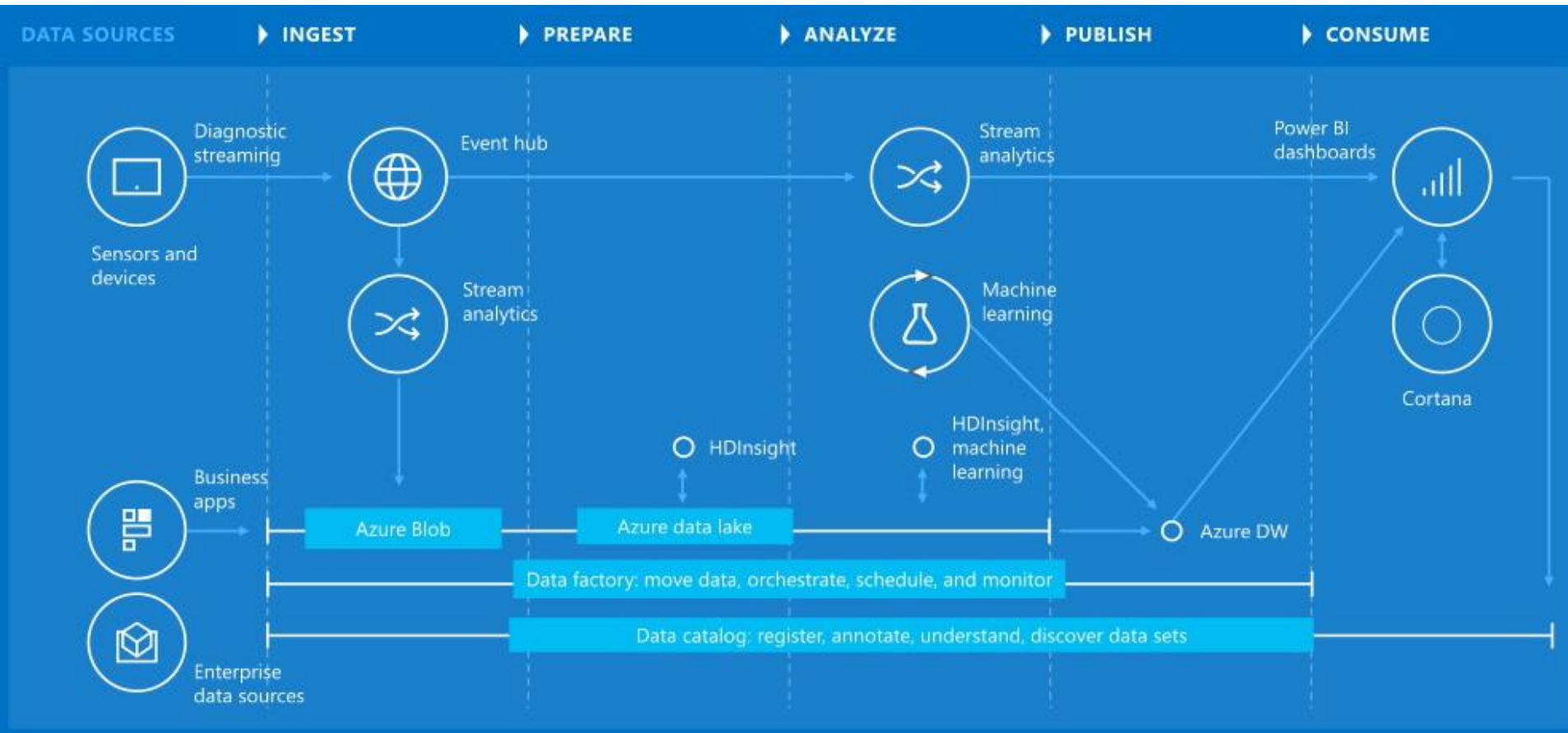
Permutation Feature Importa... ✓

Score Model ✓

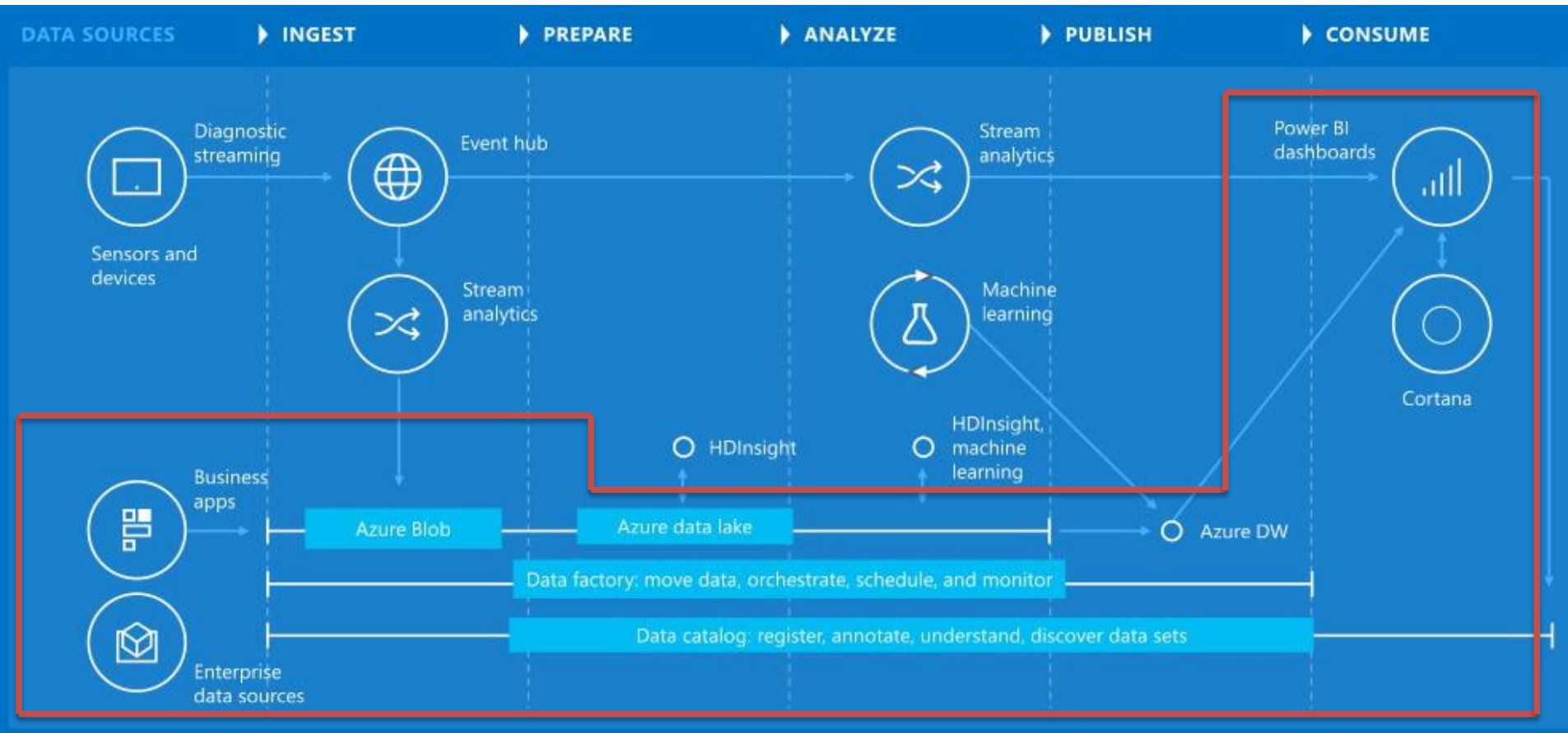
Evaluate Model ✓

NEW | VIEW RUN HISTORY | SAVE | SAVE AS | DISCARD CHANGES | REFRESH | CANCEL | RUN | PREPARE WEB SERVICE | PUBLISH TO GALLERY | UPDATE SCORING EXPERIMENT

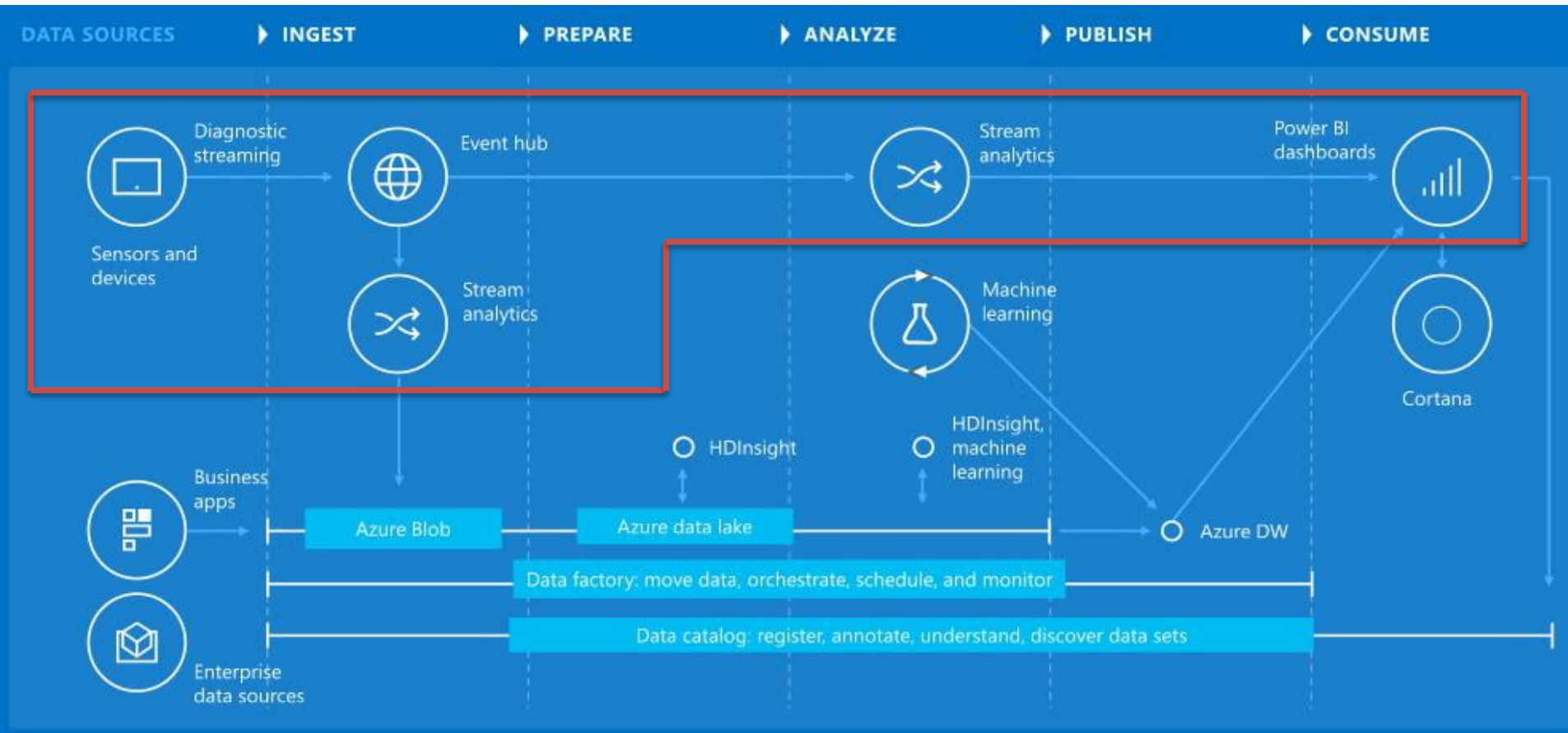
Cortana Intelligence



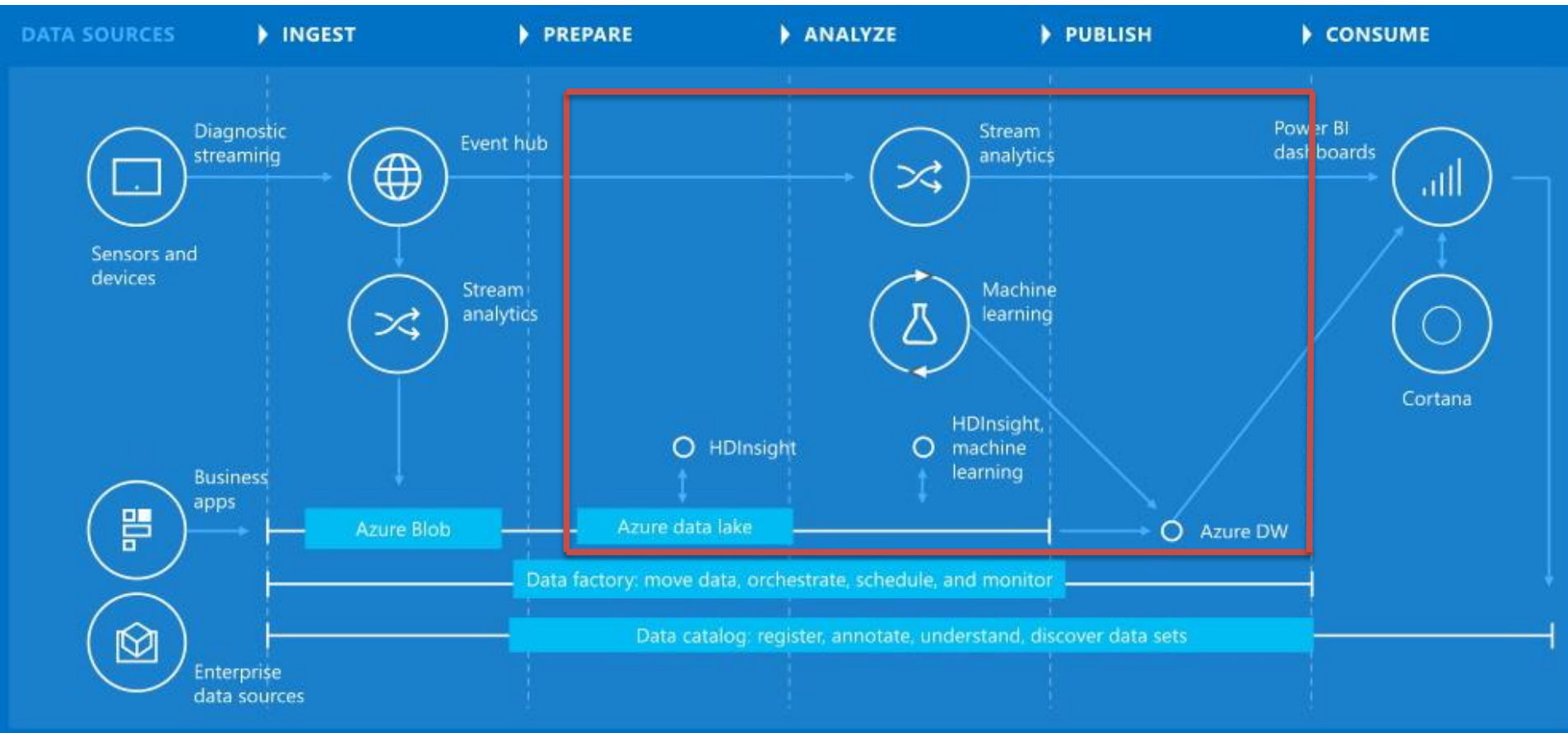
Cortana Intelligence - Batch



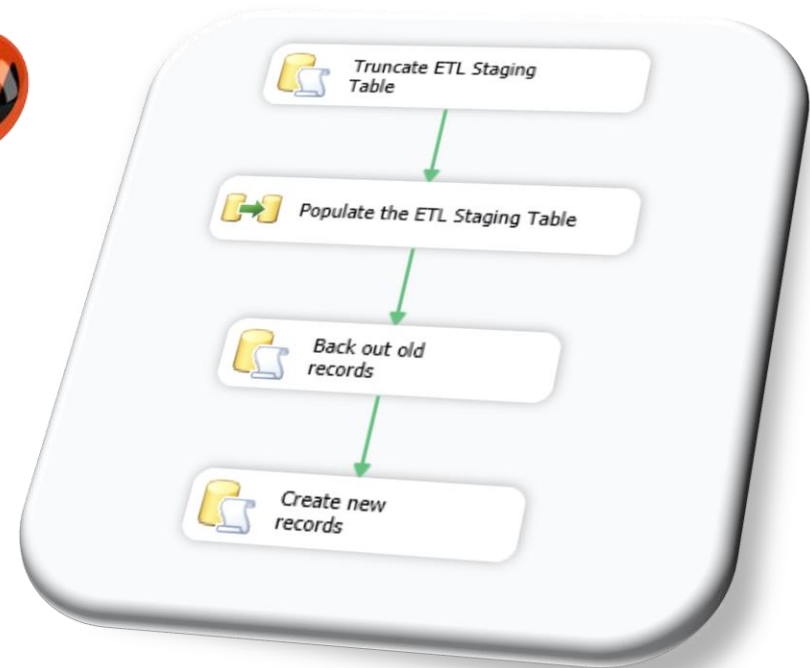
Cortana Intelligence - RealTime



Cortana Intelligence – Analytics



U-SQL / Data Lake Demo



U-SQL

```
DECLARE @INPUT_FILE string = @"/landing/{fileyear}/{filemonth}/{file}" ;
```

```
@rawdata =
```

```
  EXTRACT dt string
```

```
    ,ip string
```

```
    ,address string
```

```
    ,gender string
```

```
    ,logtext string
```

```
    ,fileyear int
```

```
    ,filemonth int
```

```
    ,file string
```

```
FROM @INPUT_FILE
```

```
USING Extractors.Text(delimiter:' ');
```

** Schema on read*

U-SQL

```
@processeddata =  
    SELECT dt.Substring(5) AS dt  
        ,ip  
        ,((gender== "M" || gender== "F") ? gender : "") AS gender  
        ,logtext.Substring(11).Replace("]", "").Replace(":", "") AS logtext  
        ,MyProject.udf.ExtractHouseNumber("address") AS housenumber  
  
        ,ROW_NUMBER() OVER (PARTITION BY ip ORDER BY dt) AS rn  
FROM @rawdata  
WHERE fileyear == 2017;  
  
OUTPUT @processeddata  
TO "/out/mydata.csv"  
USING Outputters.Text(quoting : true,  outputHeader : true);
```

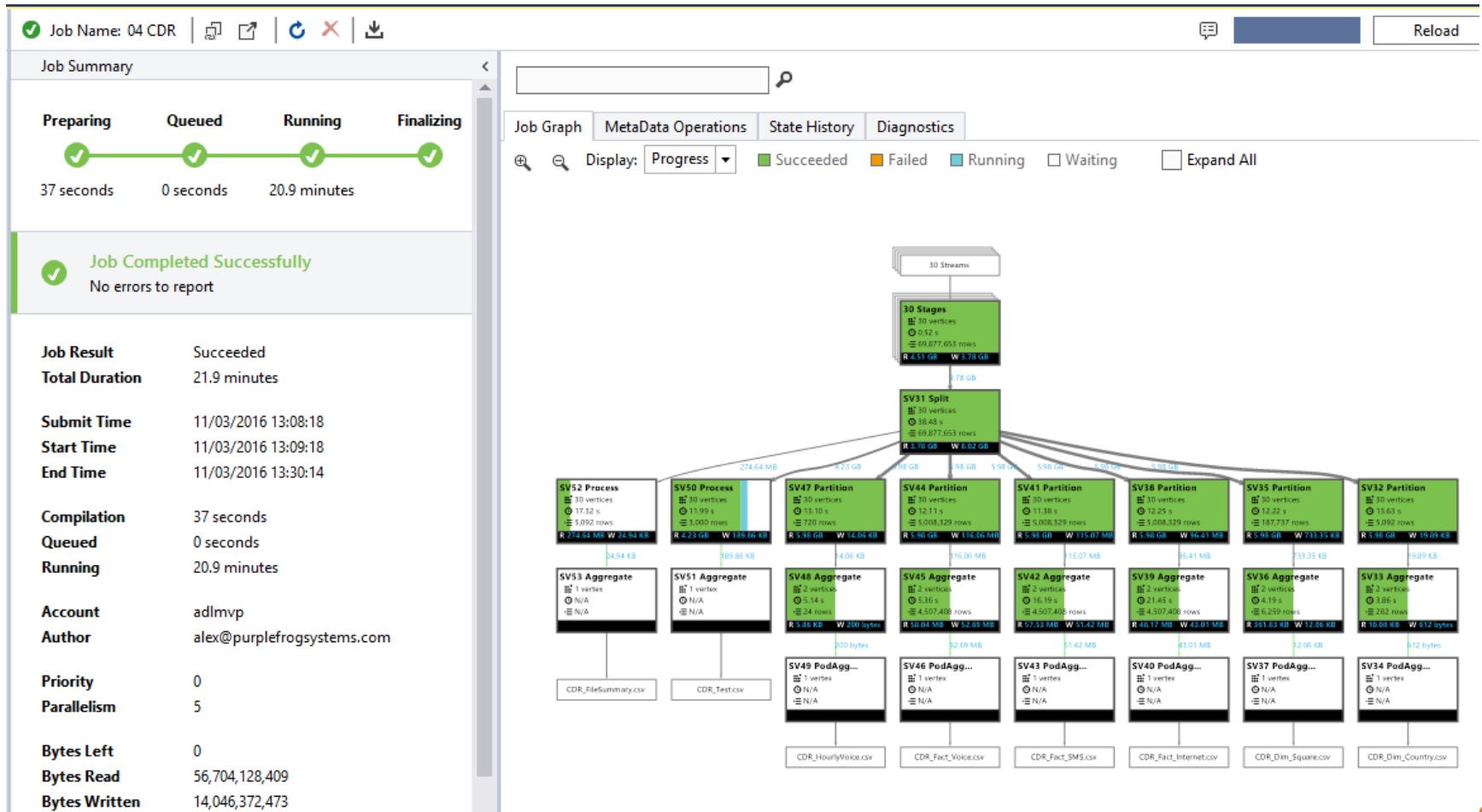
U-SQL – User Defined Functions

```
static public string ExtractHouseNumber(string Address1)
{
    //Pattern: Start of string, any integers, 0 or 1 letter, end of word
    string sPattern = @"^[0-9]+([A-Za-z]\b)?";

    //Find any matches of the pattern in the string
    Match match = Regex.Match(Address1, sPattern, RegexOptions.IgnoreCase);

    //If a match is found
    if (match.Success)
    {
        //Return the first match into the new
        //HouseNumber field
        return match.Groups[0].Value;
    }
    else
    {
        //If not found, leave the HouseNumber blank
        return "";
    }
}
```

U-SQL / Data Lake



U-SQL / Data Lake

Job Browser: adlmvp - Microsoft Visual Studio

File Edit View Project Debug Team Data Lake Architecture Test SQL Prompt Tools Analyze Win

Server Explorer

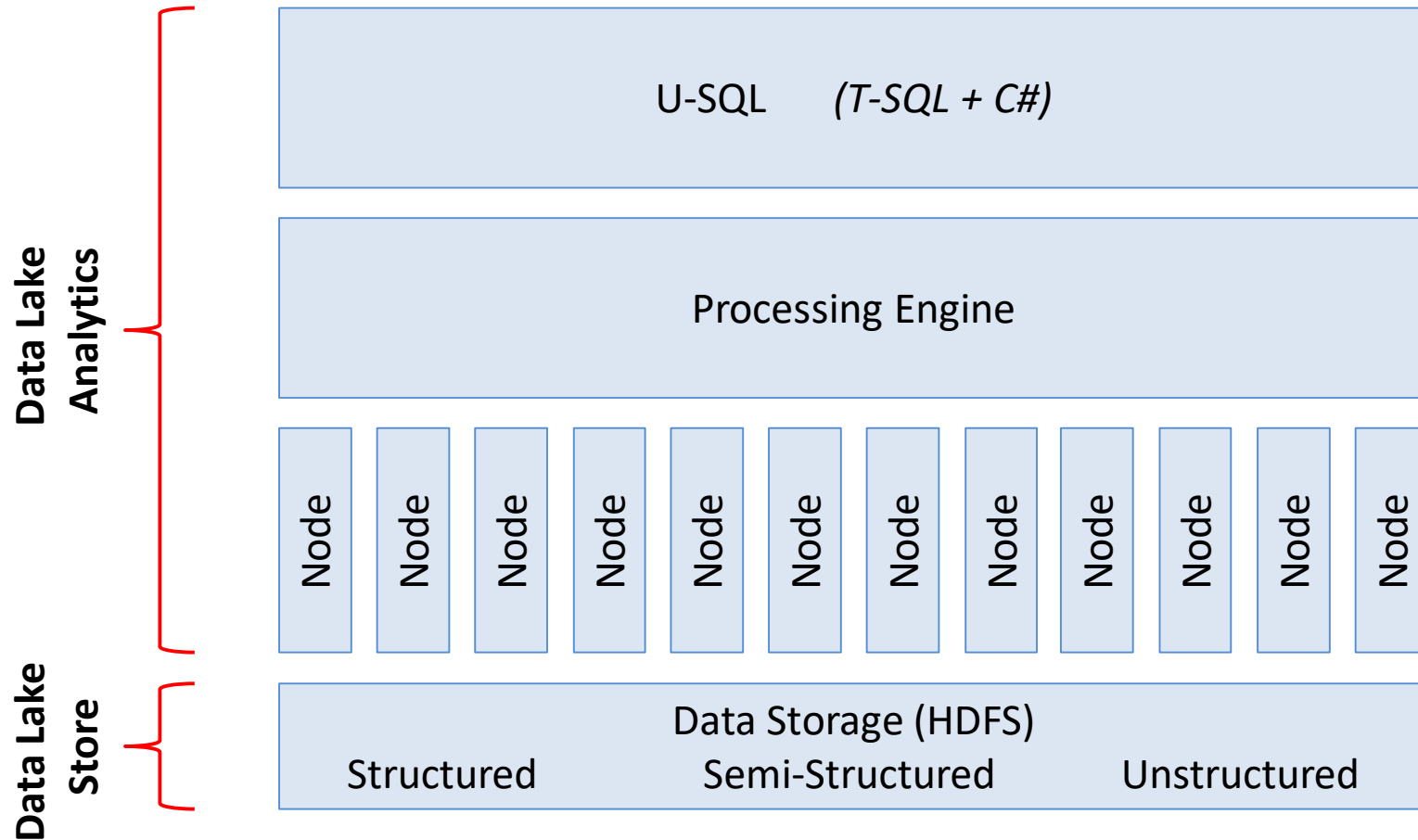
- Azure (alex@purplefrogsystems.com -)
 - App Service
 - Cloud Services
 - Data Factory
 - Data Lake Analytics
 - (Local)
 - adlmvp
 - Databases
 - Damian_SampleDB
 - DriverDB
 - InfoMeetKatowice
 - InfoMeetKrakow
 - MarketDataTest
 - SQLPort
 - SampleDBTutorials
 - SampleDBTutorialsLuk
 - Triangles
 - master
 - Jobs
 - Storage Accounts
 - adlmvp(Default Storage
 - HDInsight
 - Mobile Services
 - Notification Hubs
 - Service Bus
 - SQL Databases
 - Storage
 - Virtual Machines
 - Data Connections
 - Servers

Job Browser: adlmvp Azure Data Lake Explorer: adlmvp

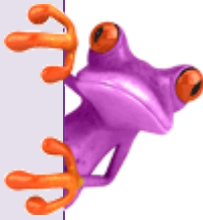
My All

	Name	Submitter
✗	1-SimpleAnalysis-1File	damar123@wp.pl
✗	1-SimpleAnalysis-1File	damar123@wp.pl
✓	Query a TSV file	damar123@wp.pl
✓	Identify unsuccessful requests	w.strasser@cubido
✓	Query [SampleDBTutorials].[dbo].[ReferrersPerDay]	w.strasser@cubido
✓	Analyze referrers	w.strasser@cubido
⊘	Analyze referrers	w.strasser@cubido
✓	Create Weblogs TVF	w.strasser@cubido
✓	Query a TSV file	w.strasser@cubido
✓	01 Twitter	alex@purplefrogsy
✓	01 Twitter	alex@purplefrogsy
✓	04 CDR	alex@purplefrogsy
✓	02 Twitter	alex@purplefrogsy
✓	01 Twitter	alex@purplefrogsy
✓	03 Script	alex@purplefrogsy
✓	02 Twitter	alex@purplefrogsy
✓	02 Twitter	alex@purplefrogsy
✓	03 Script	alex@purplefrogsy
✓	01 Twitter	alex@purplefrogsy
✓	Query [SampleDBTutorials].[dbo].[ReferrersPerDay]	blackzero@hotmail
✓	2. Ordering_Grouping_Aggregation	info@webcaravela
✓	1. Basics	info@webcaravela
✓	4. Database_structures	info@webcaravela
✓	Luk TestCDR p1	lqrala@o2.pl

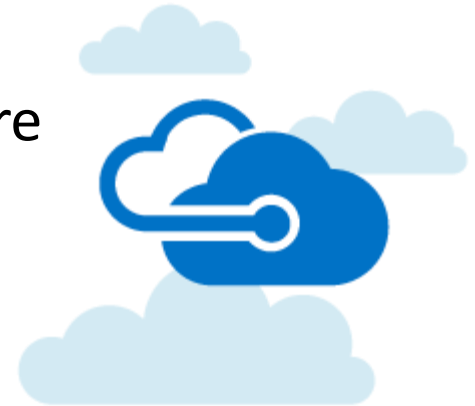
Data Lake



In Summary...



- Virtual Machine with On-Prem solution?
- Or genuine Azure solution?
- Real-time, Batch, or both? Lambda Architecture
- U-SQL & Data Lake <- Everything to Everyone!
- E C T L T ?
- Missing:
 - SSIS Data flow equivalent



Business Intelligence in Azure



Alex@PurpleFrogSystems.com

[@PurpleFrogSys](#)

www.PurpleFrogSystems.com

www.PurpleFrogSystems.com/blog