Business Intelligence in Azure

SQL GRILLEN

Alex Whittles

purple frog
business intelligence consultancy

Alex@PurpleFrogSystems.com

PurpleFrogSystems.com

PurpleFrogSystems.com/blog

@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 purplefrog BI Consultancy

MVP
Microsoft®
Most Valuable Professional

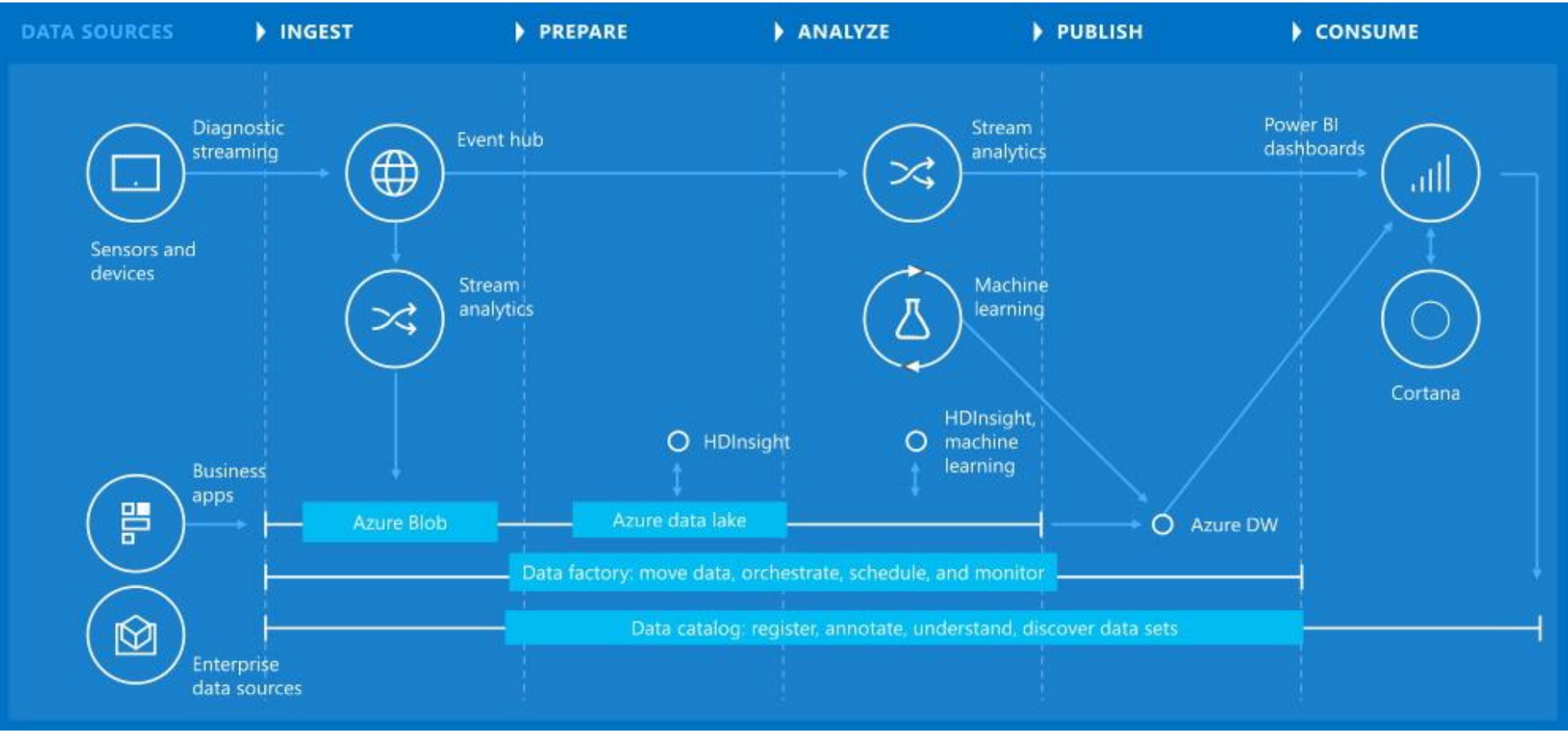
Business Intelligence Consultancy

- Data Modelling
- Data Warehousing
- OLAP Cubes
- ETL Systems
- Reporting Systems
- Managed Service

Alex Whittles

MVP
Microsoft®
Most Valuable
Professional

Cortana Analytics Data Flow



Business Intelligence in Azure

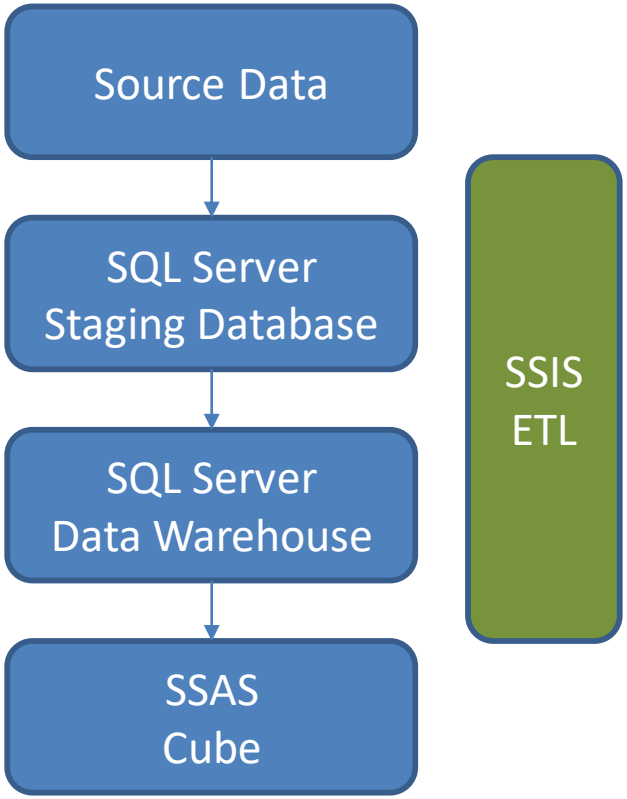
Options:

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

For each:

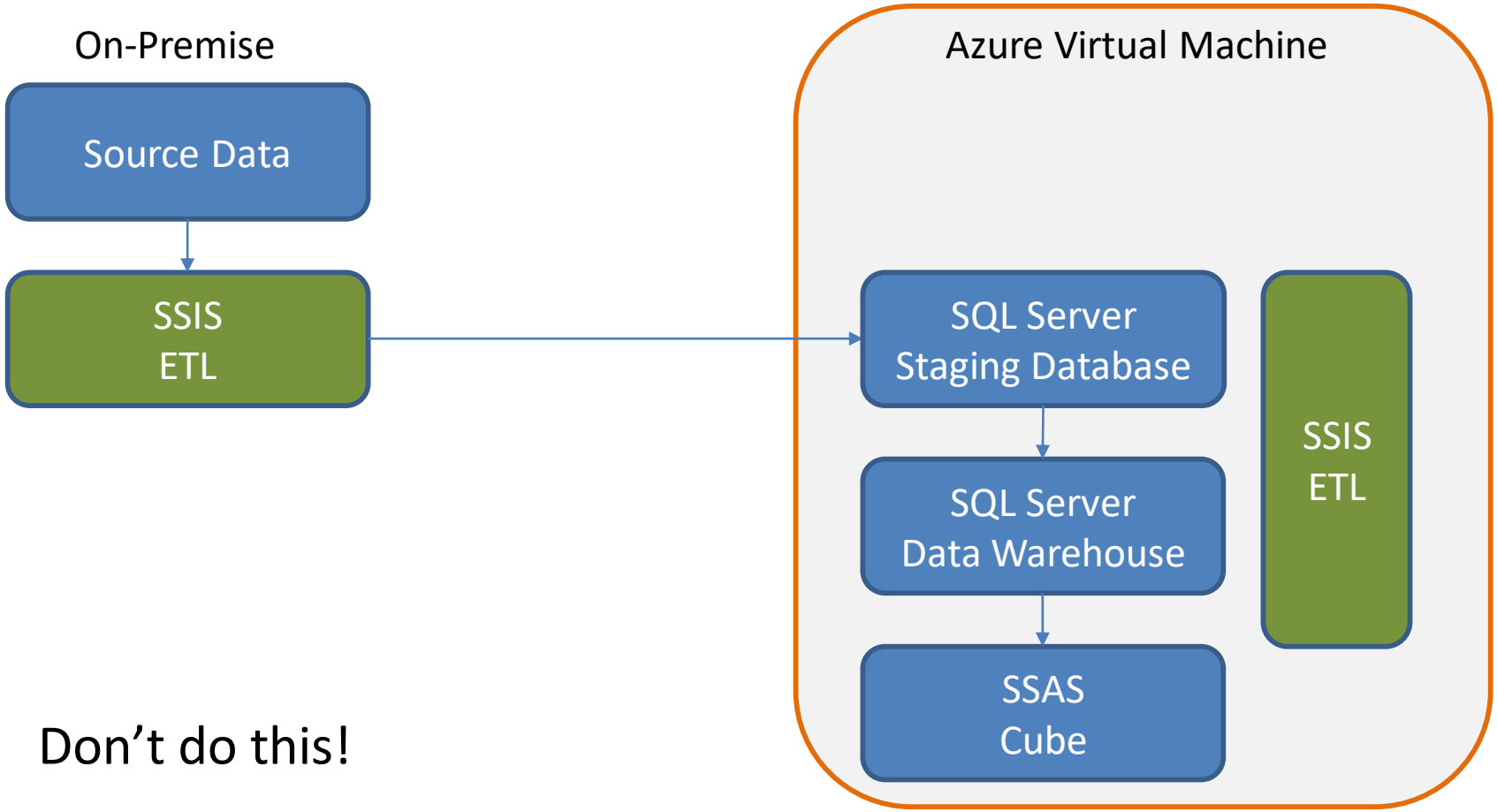
Staging / ETL / Data Warehouse / Cube

On-Premise Batch Load BI



Azure IAAS Solution

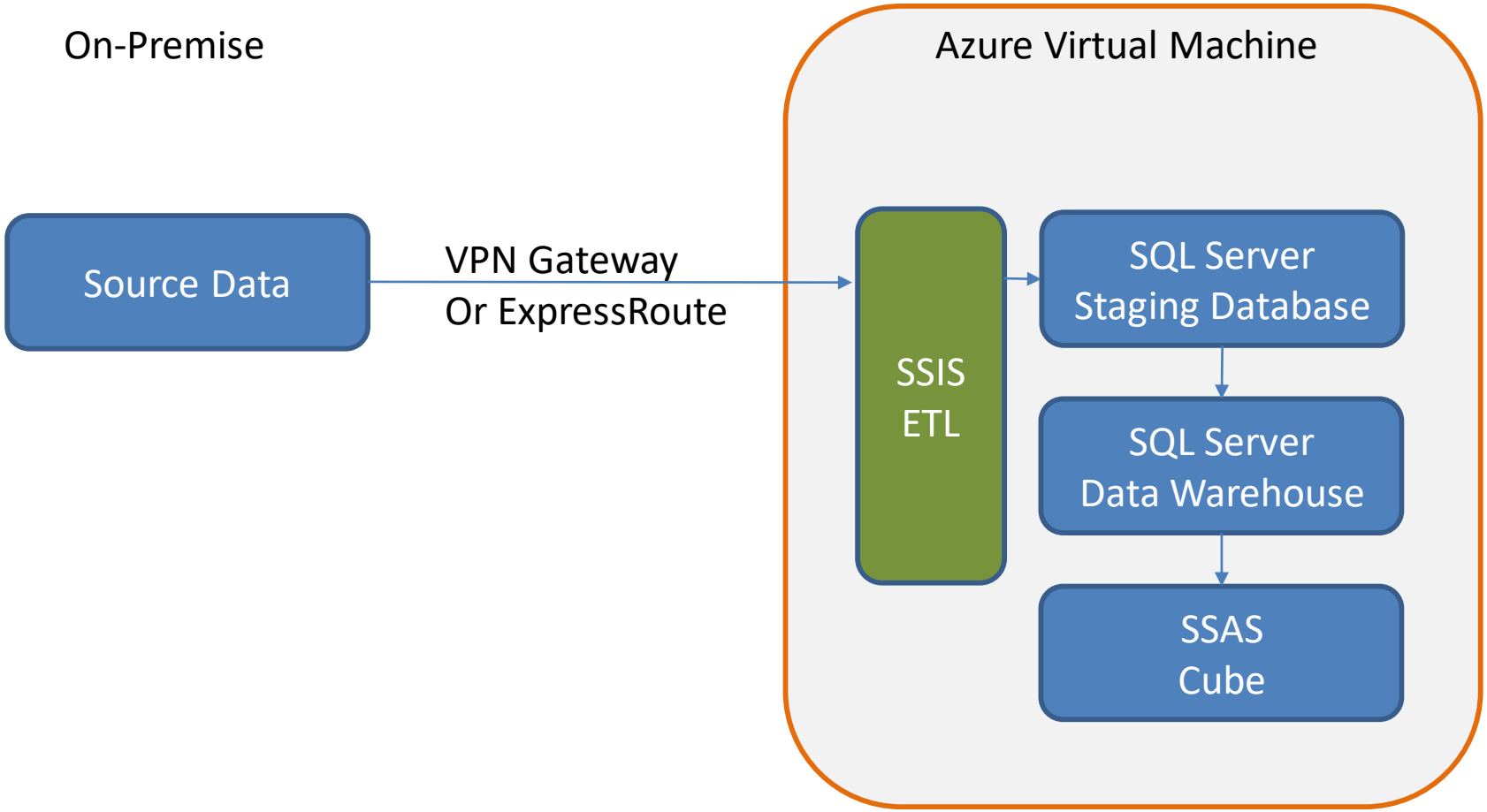
[Infrastructure As A Service]



Don’t do this!

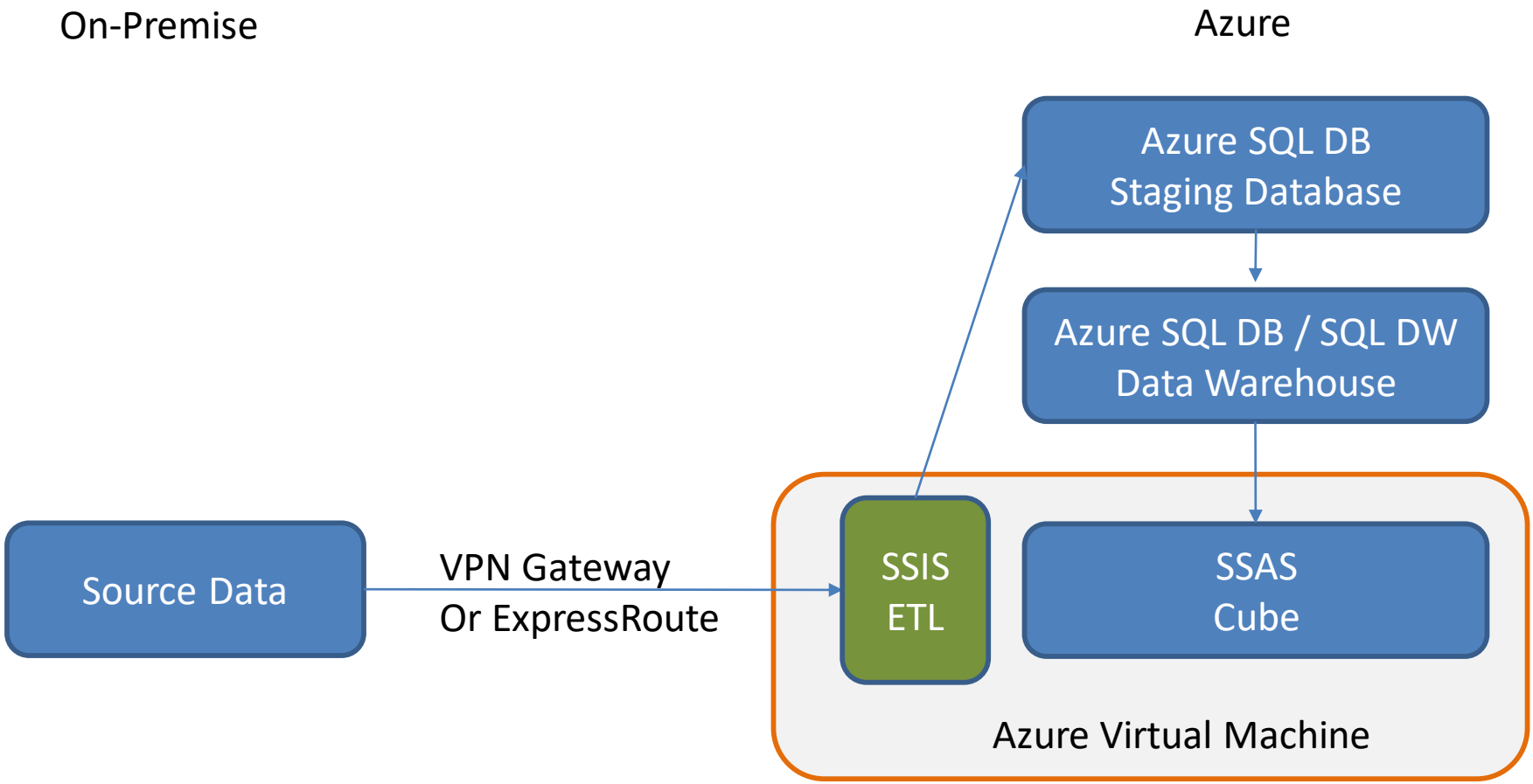
Better Azure IAAS Solution

[Infrastructure As A Service]



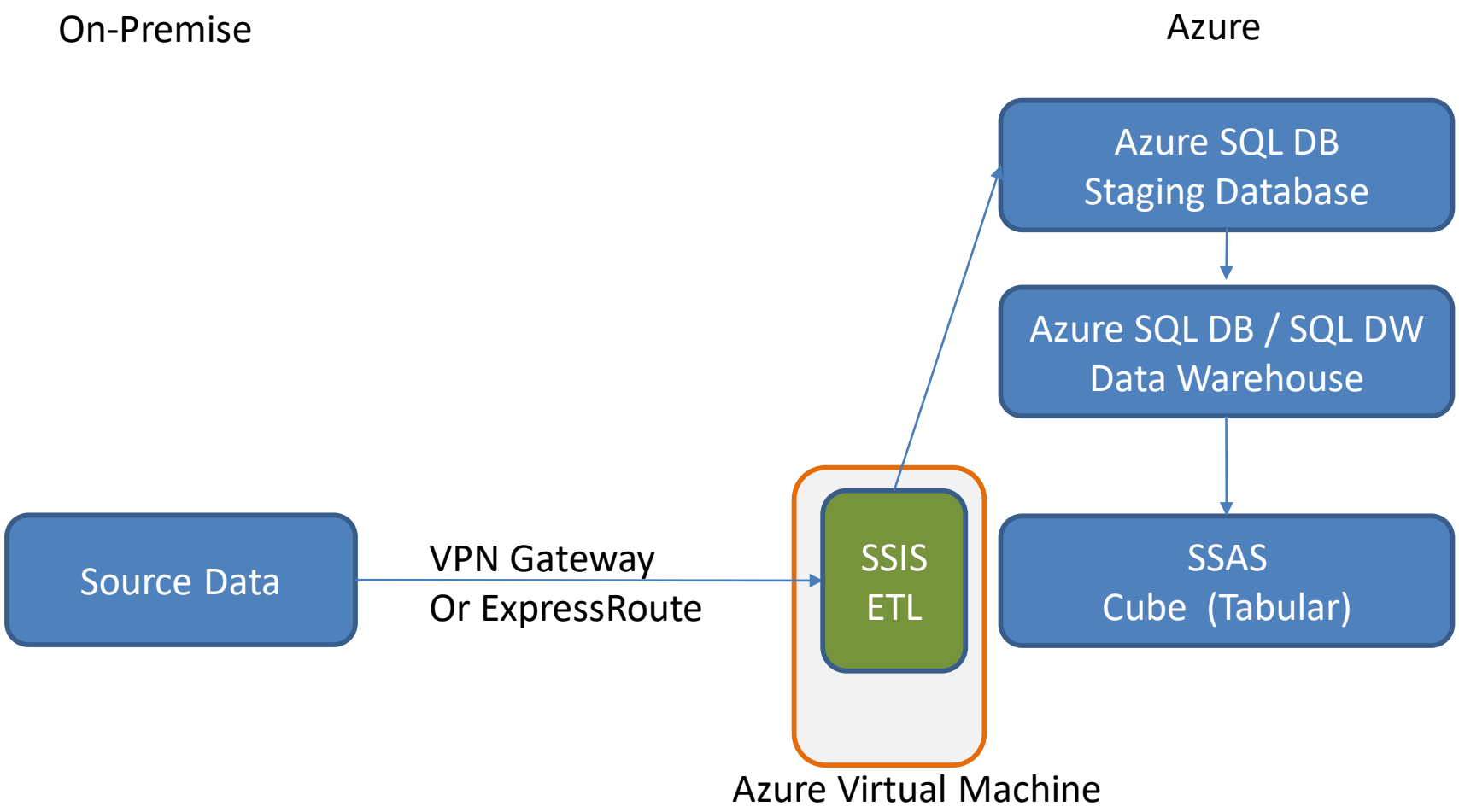
PAAS / IAAS Hybrid

[Platform As A Service]

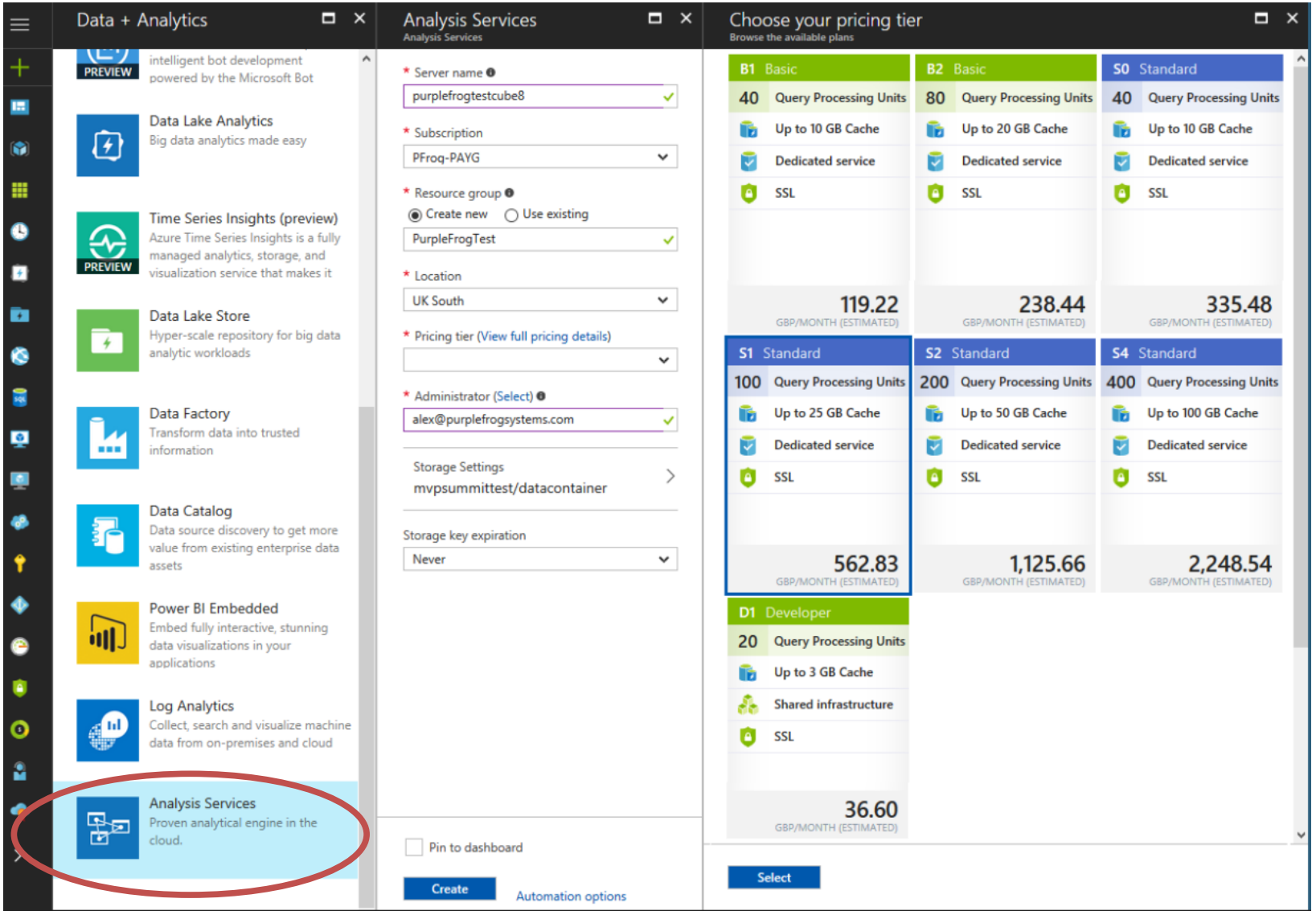


PAAS / IAAS Hybrid

[Platform As A Service]



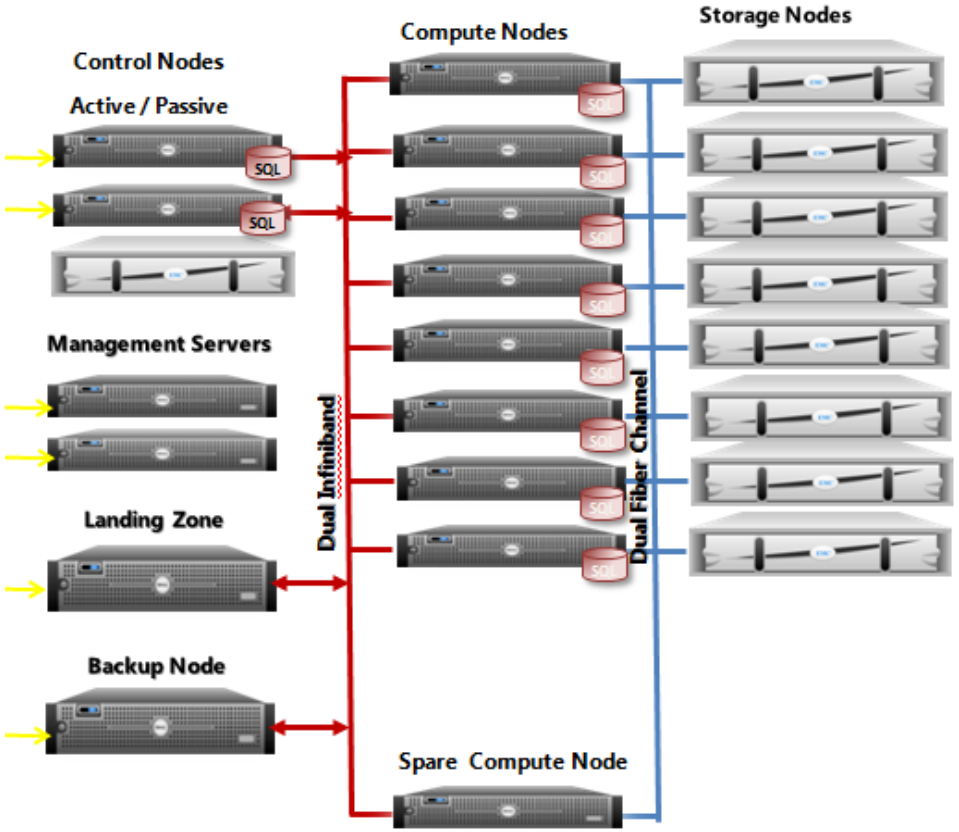
PAAS SSAS Tabular



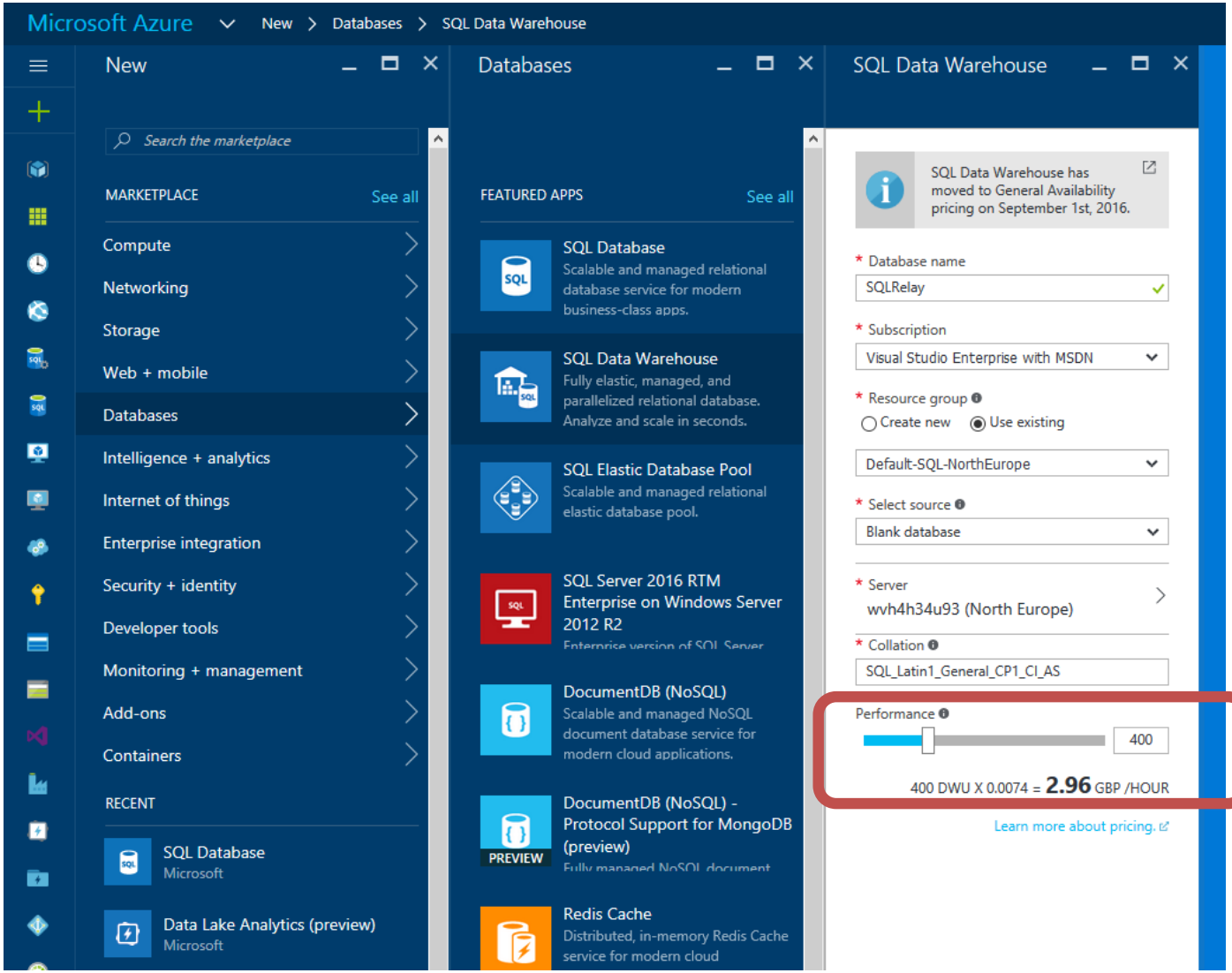
PDW / APS / SQL DW

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



Azure SQL DW – Table Distribution

Round Robin

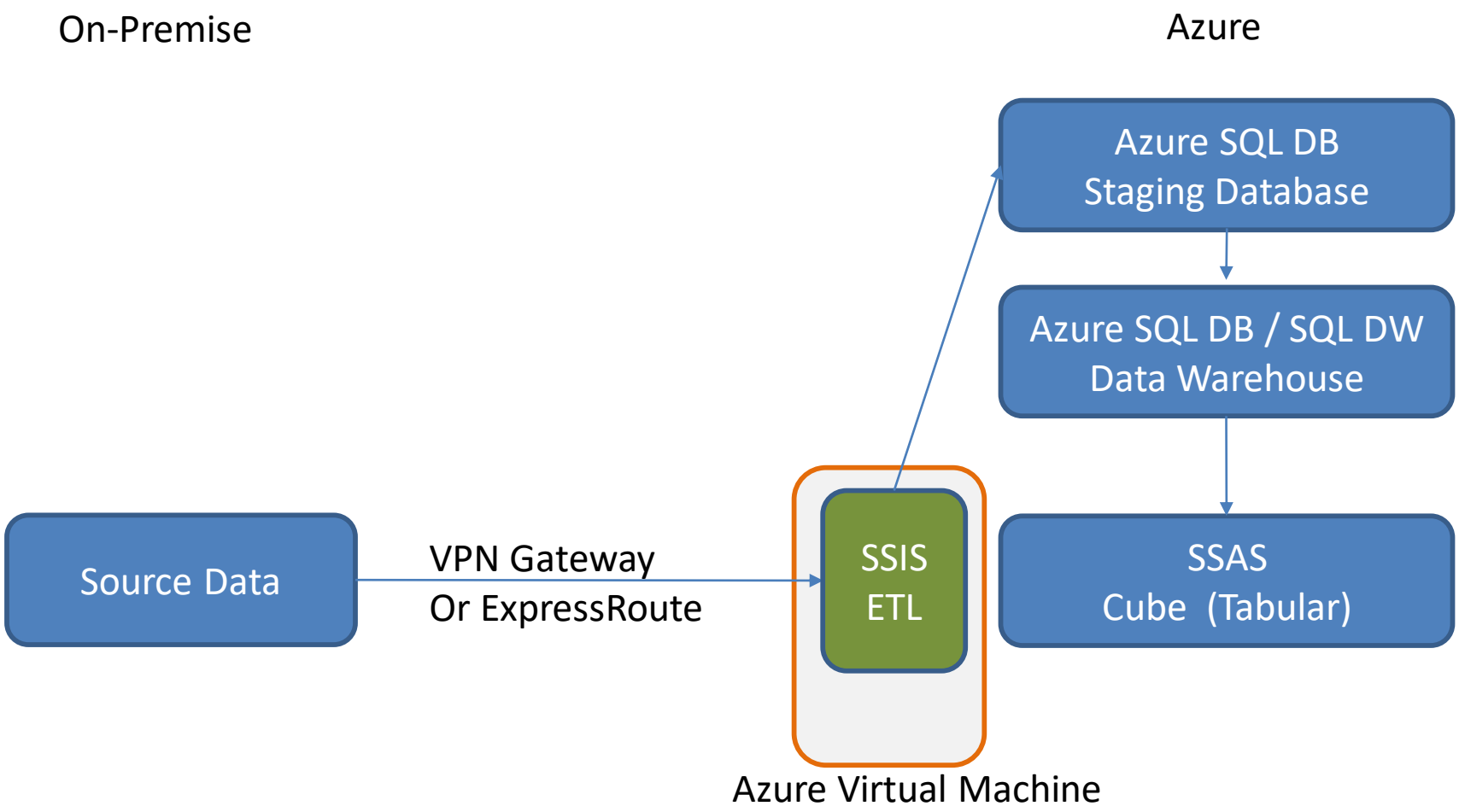
```
CREATE TABLE [dbo].[FactInternetSales]
(    [ProductKey]            int           NOT NULL
,    [OrderDateKey]          int           NOT NULL
,    [CustomerKey]           int           NOT NULL
,    [PromotionKey]          int           NOT NULL
,    [SalesOrderNumber]      nvarchar(20)  NOT NULL
,    [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])
)
;
```

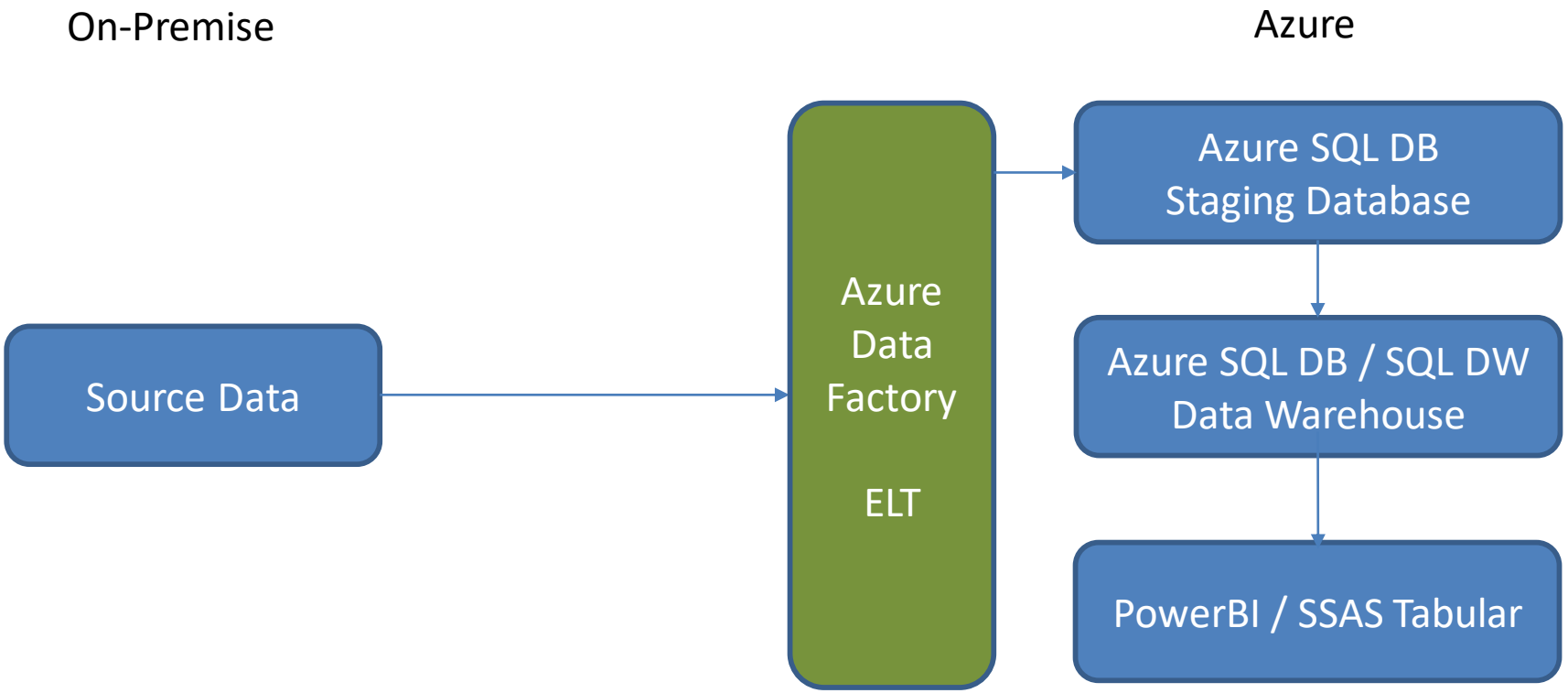
PAAS / IAAS Hybrid

[Platform As A Service]

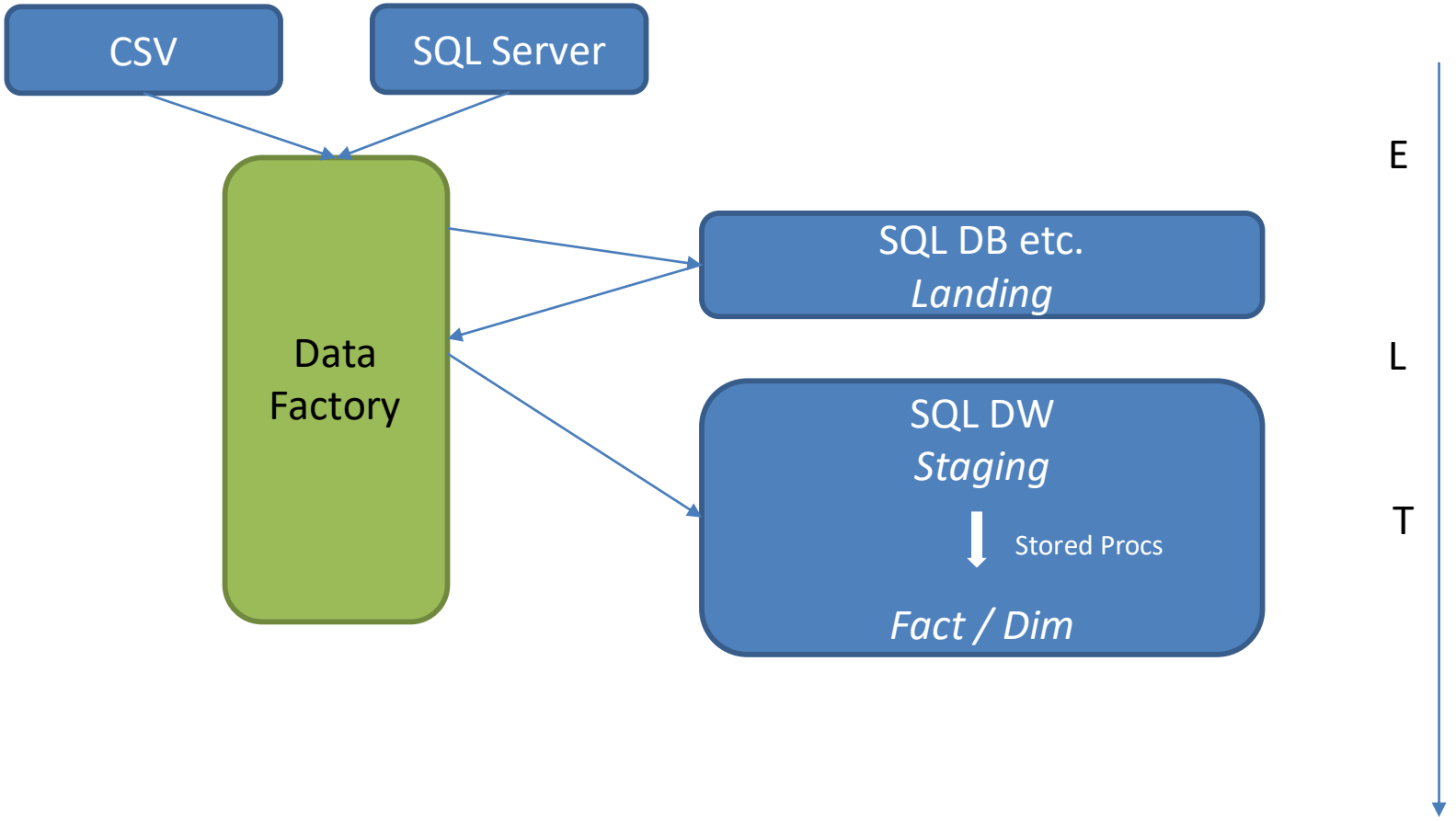


PAAS Solution

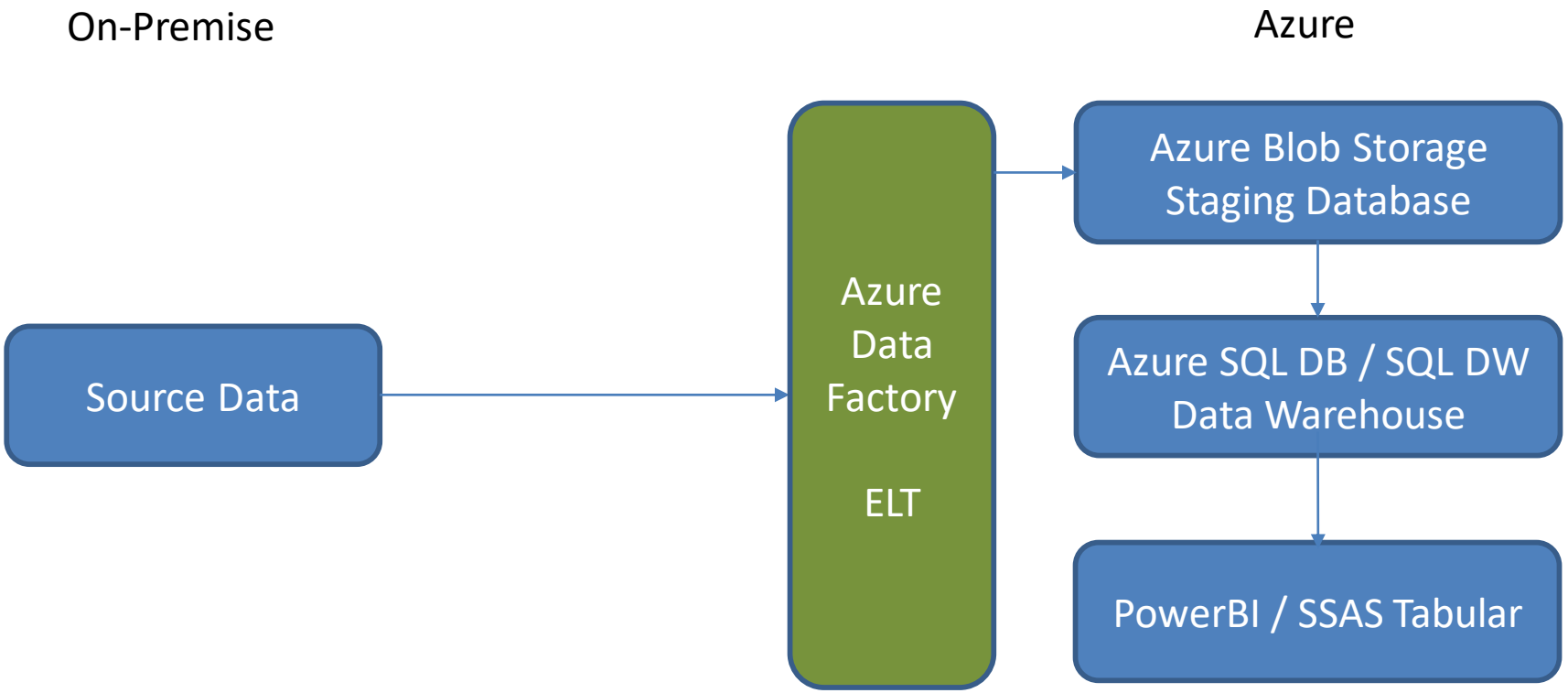
[Platform As A Service]



Example ELT Project



Cheaper Staging – Blob/Table



Blob / Table Storage

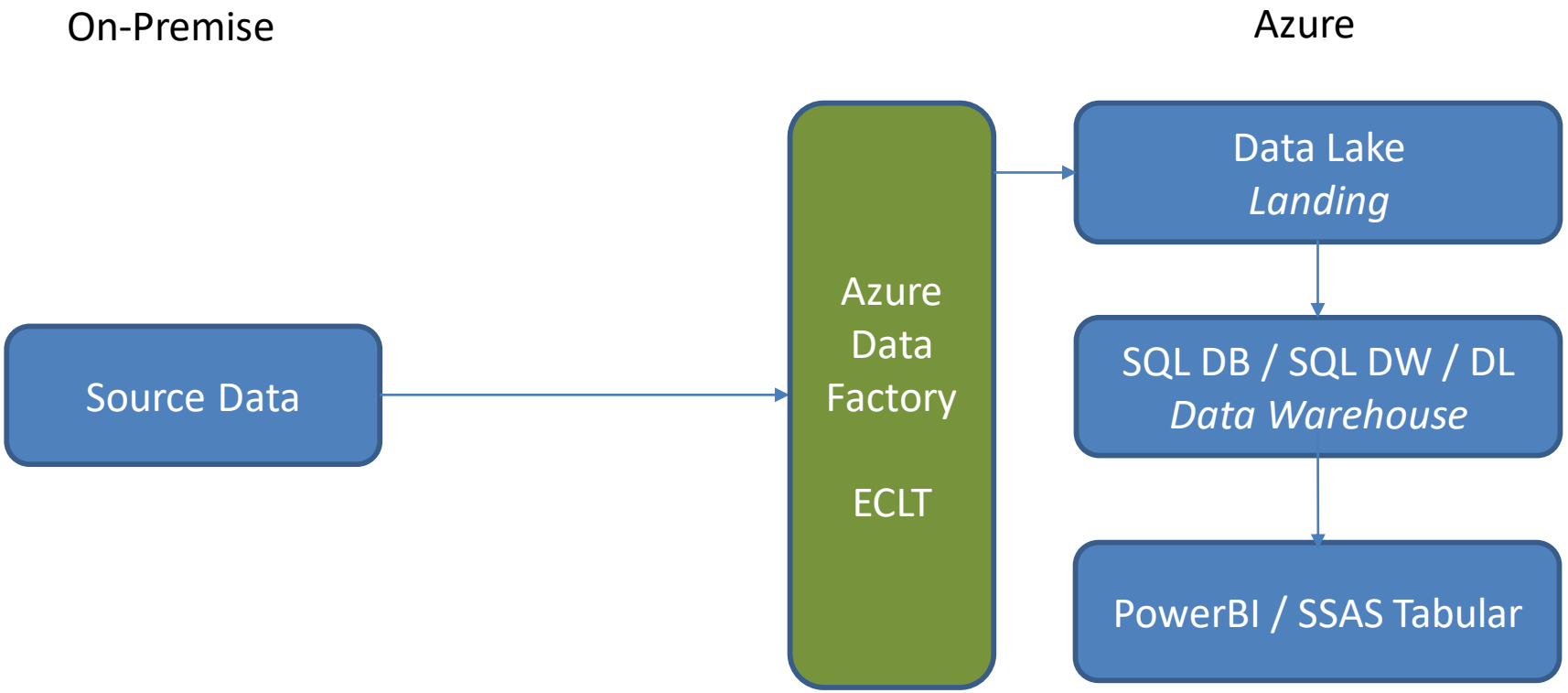
Blob

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

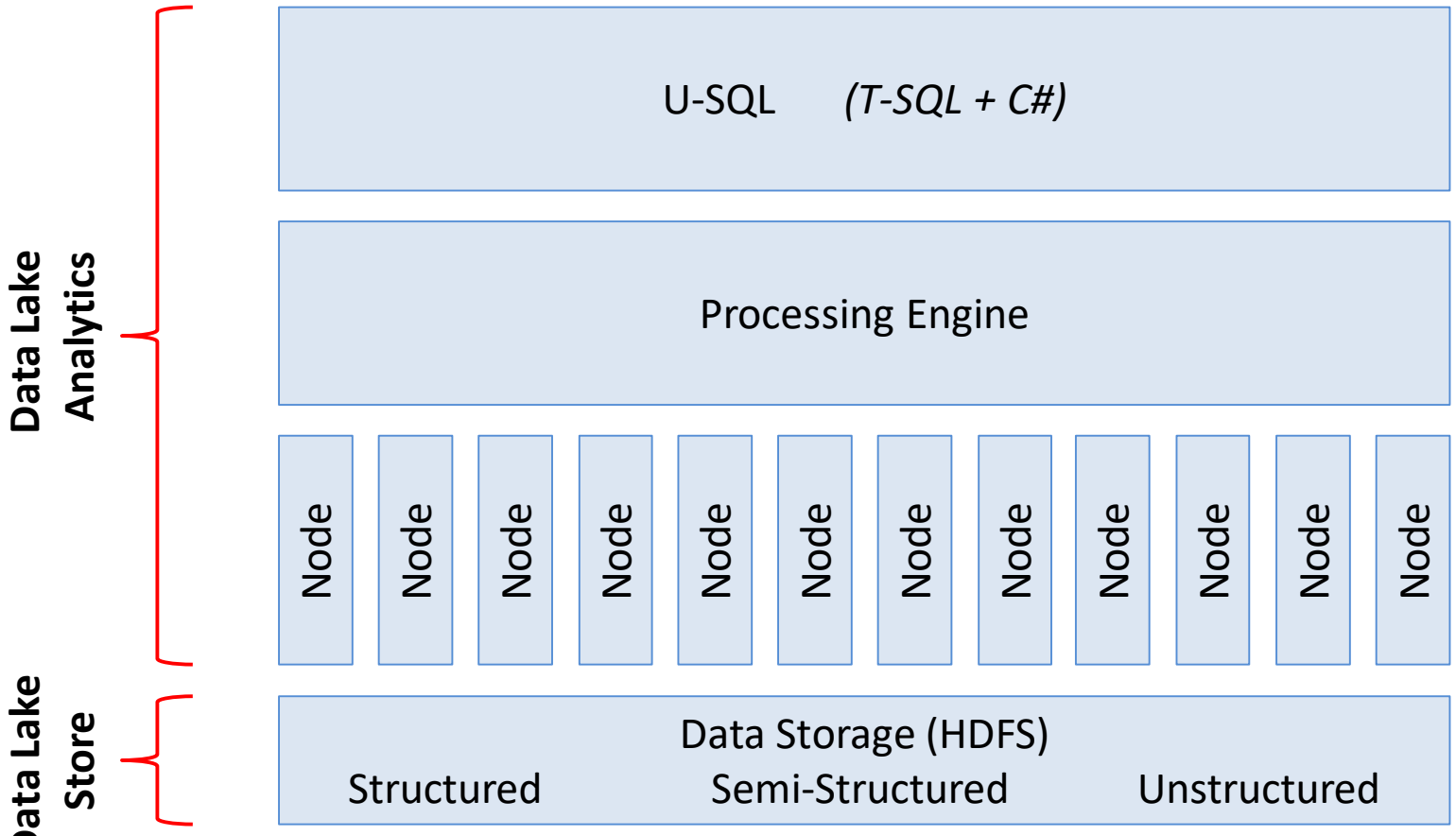
Table Storage

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

Integrate Data Lake

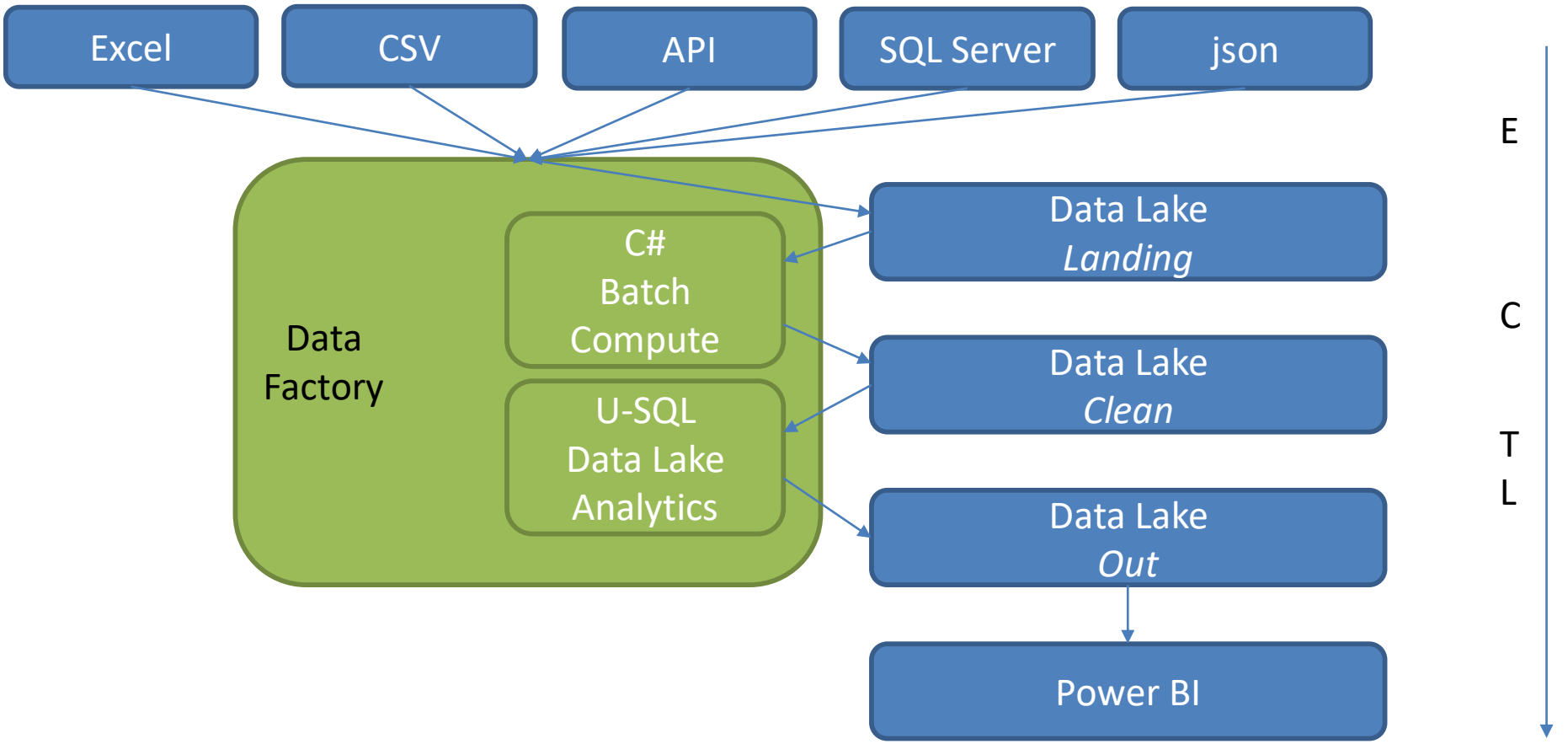


Data Lake

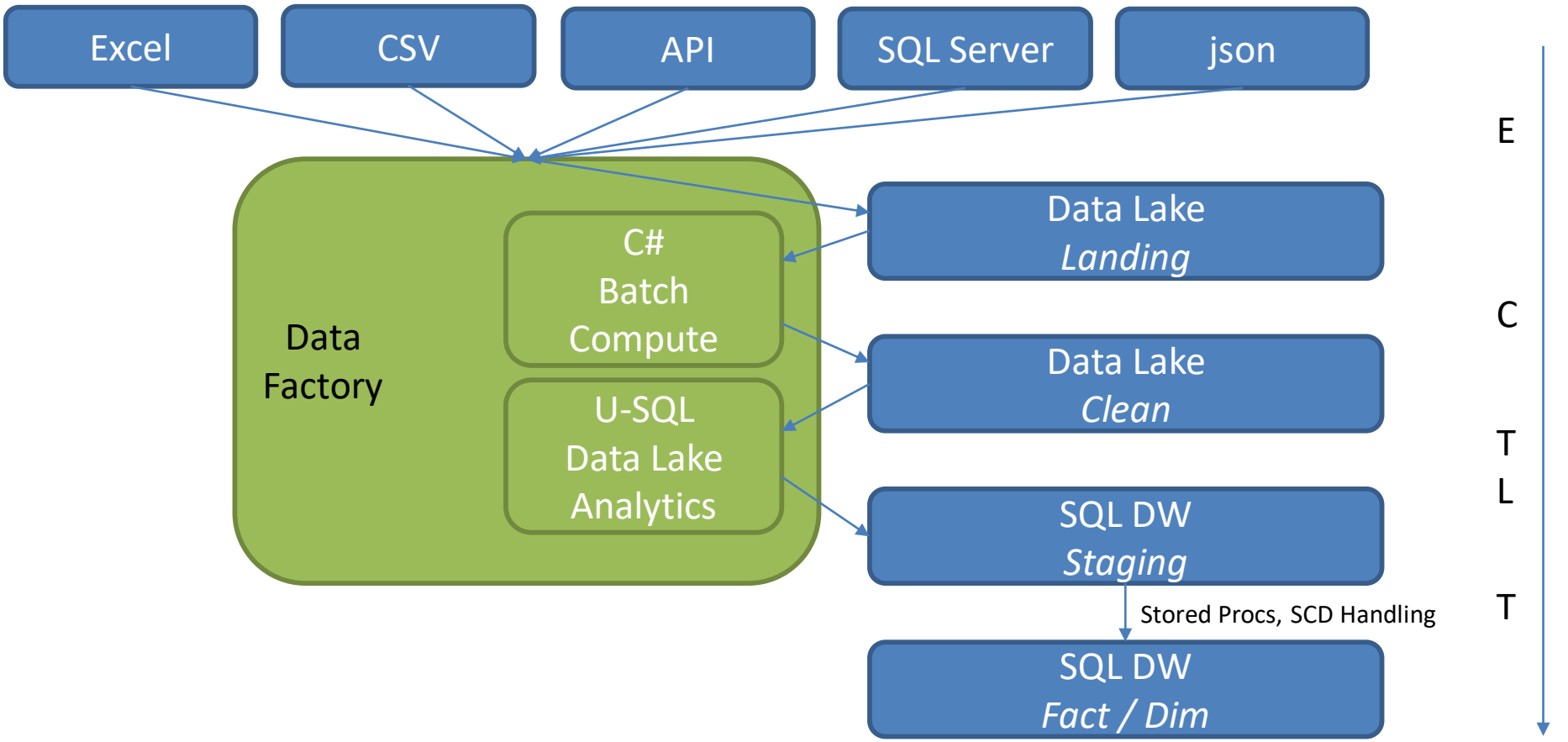


purplefrog
business intelligence consultancy

Example ECTL Project



Example ECTLT Project

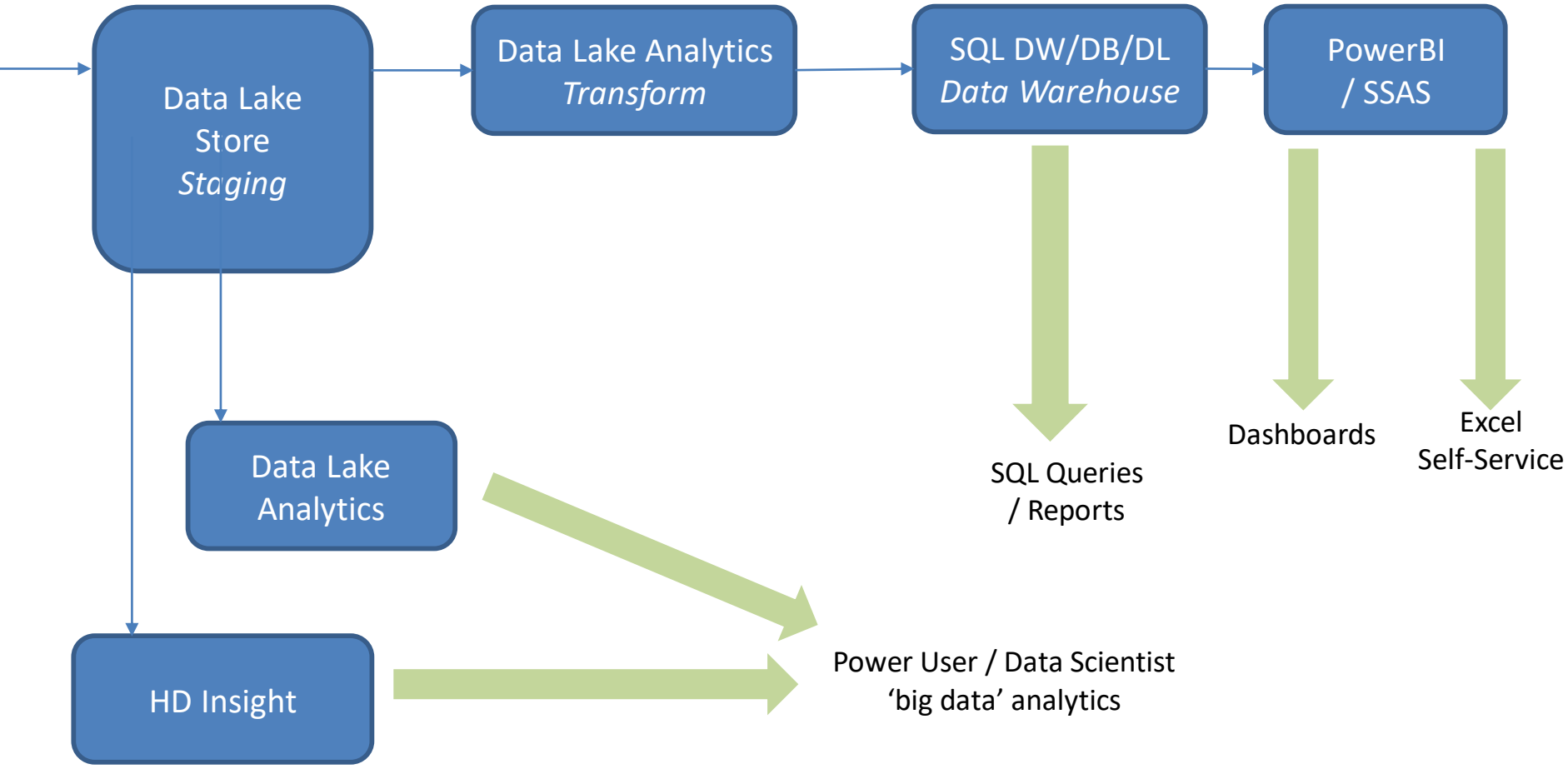


ETLT? Or ECLT? Or ECTLT?!

- 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#)

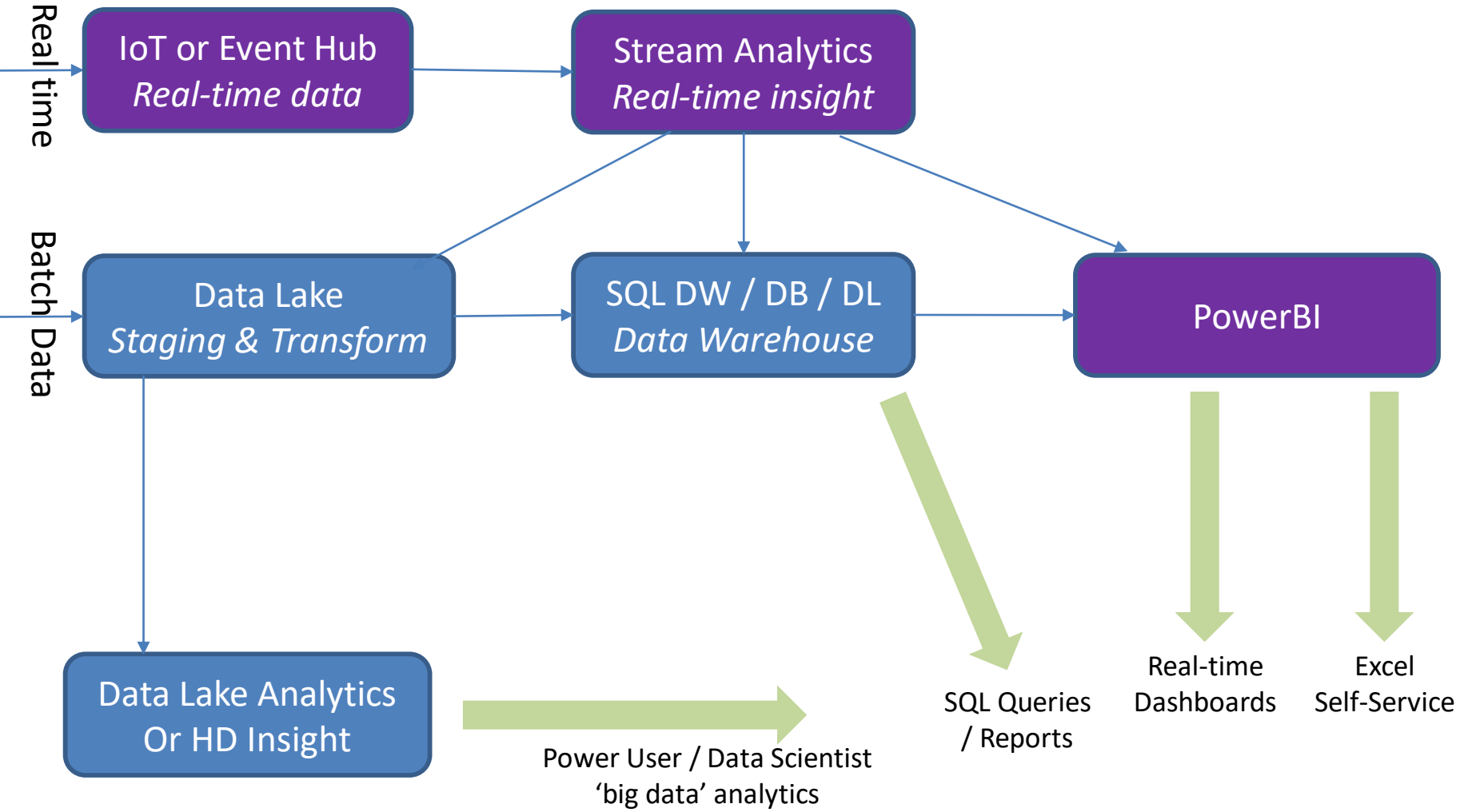
```
@CustomersFullAddress =
    SELECT CustomerName,
           Address.Split(',')[0] AS Address1,
           Address.Split(',')[1] AS Address2,
           Address.Split(',')[2] AS Address3,
           myUSQLTest.Udfs.ValidatePostCode(PostCode) AS PostCode,
           ((Gender=="M" || Gender=="F") ?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

HDInsight

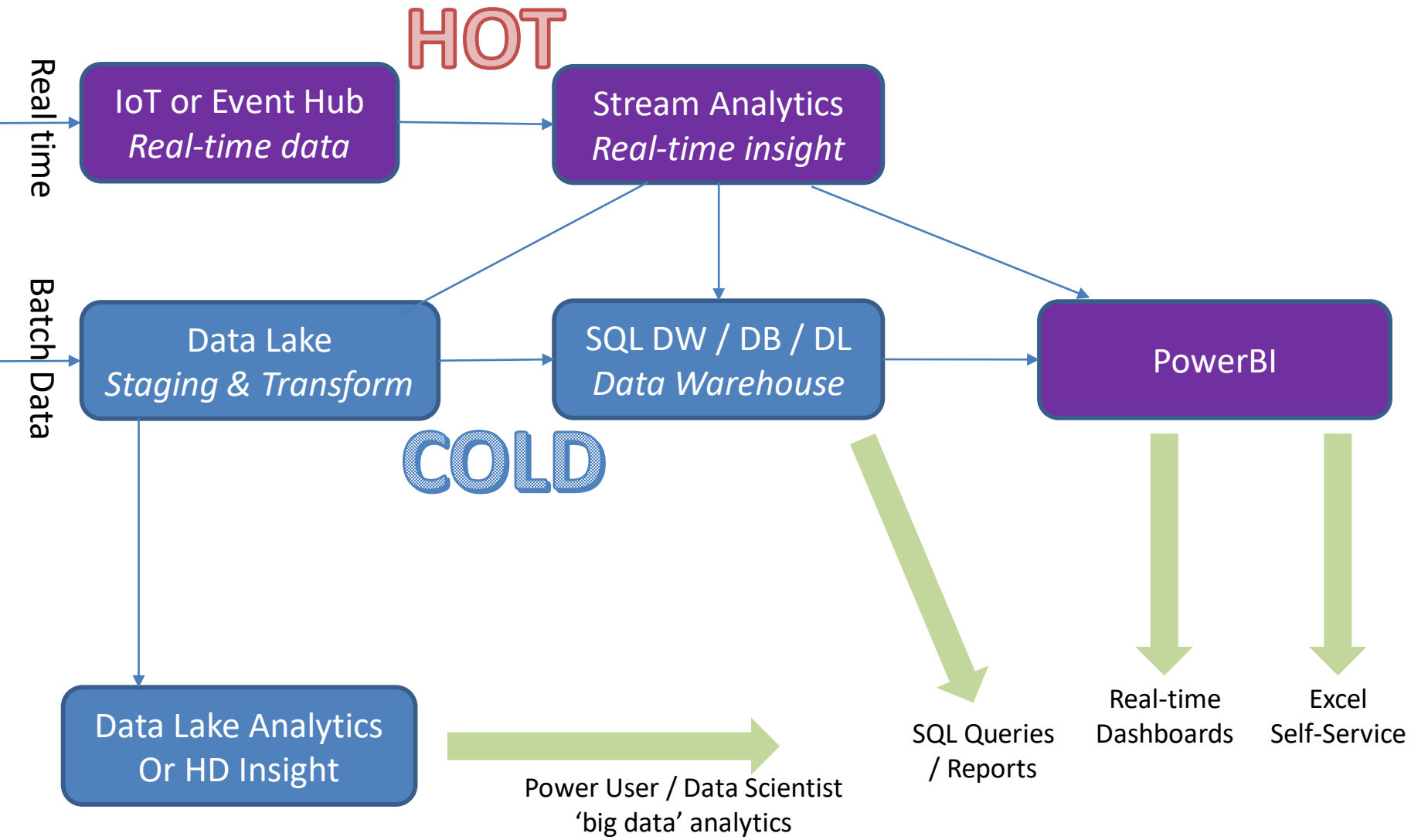
- Hadoop
- Map-Reduce
- Hive/Pig?

```
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);
  }
}
```

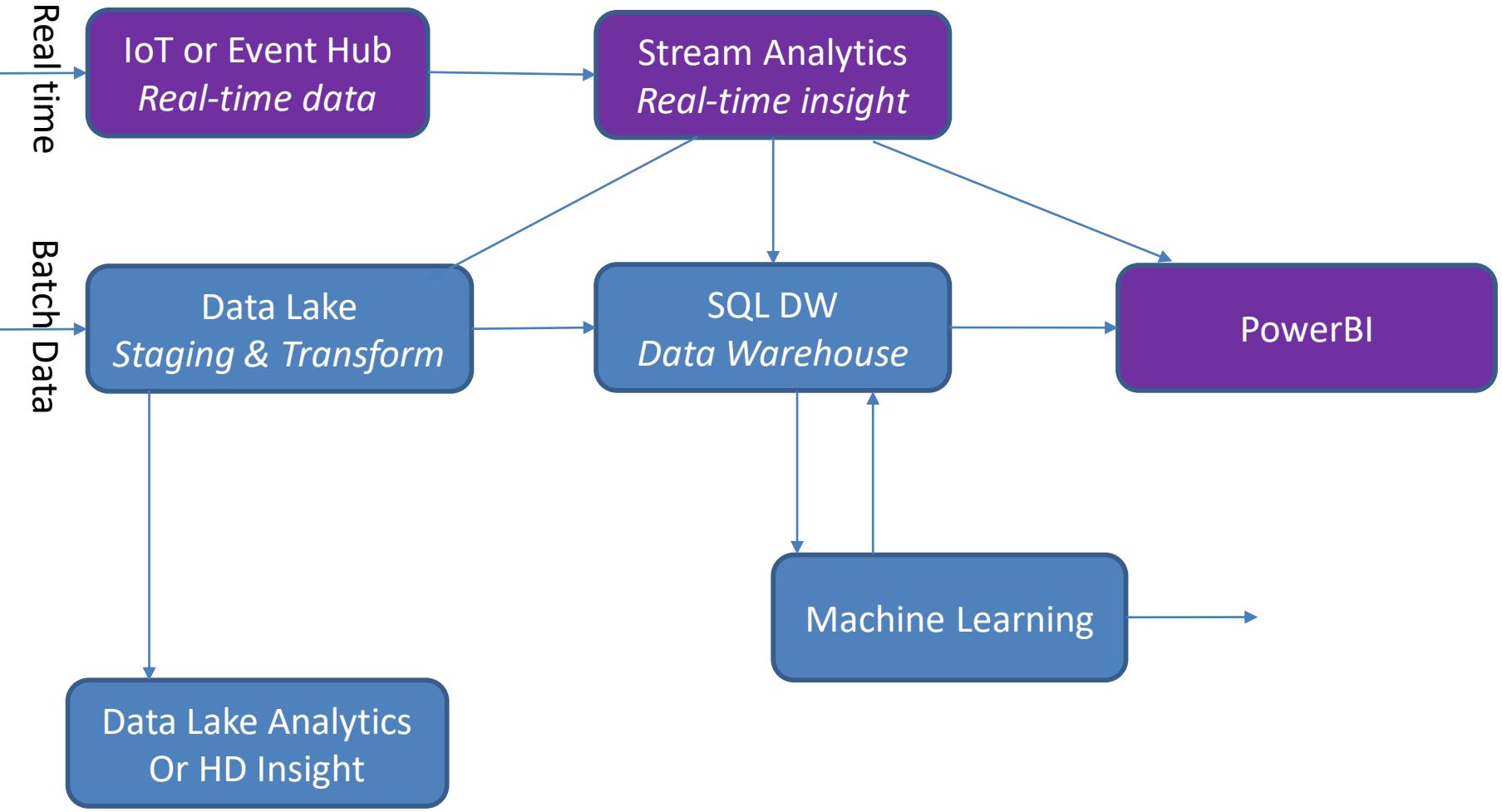
Real Time BI


Real time
Batch Data
IoT or Event Hub
Real-time data
Stream Analytics
Real-time insight
Data Lake
Staging & Transform
SQL DW / DB / DL
Data Warehouse
PowerBI
Data Lake Analytics
Or HD Insight
Power User / Data Scientist
'big data' analytics
SQL Queries
/ Reports
Real-time
Dashboards
Excel
Self-Service

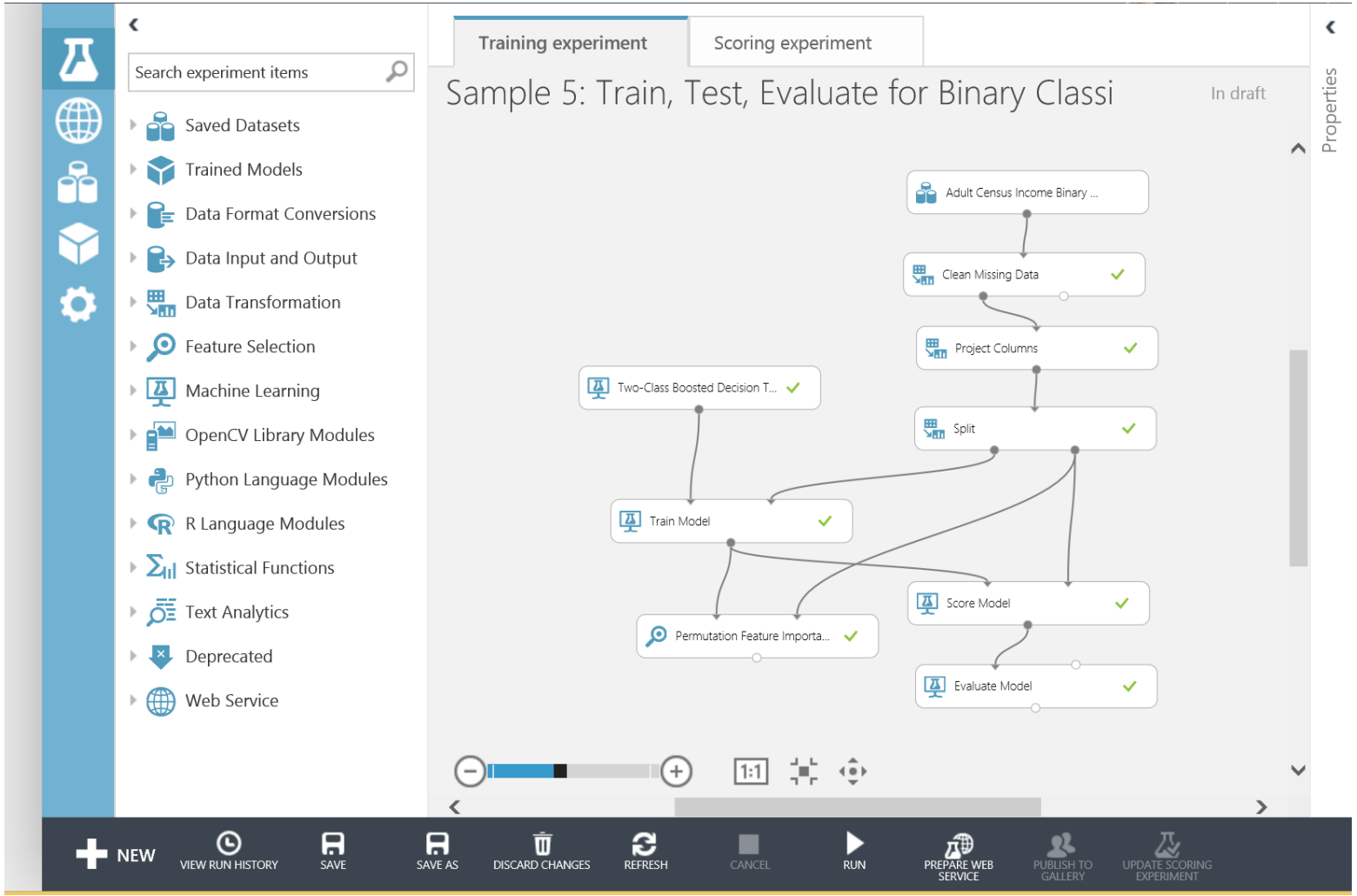
Lambda Architecture



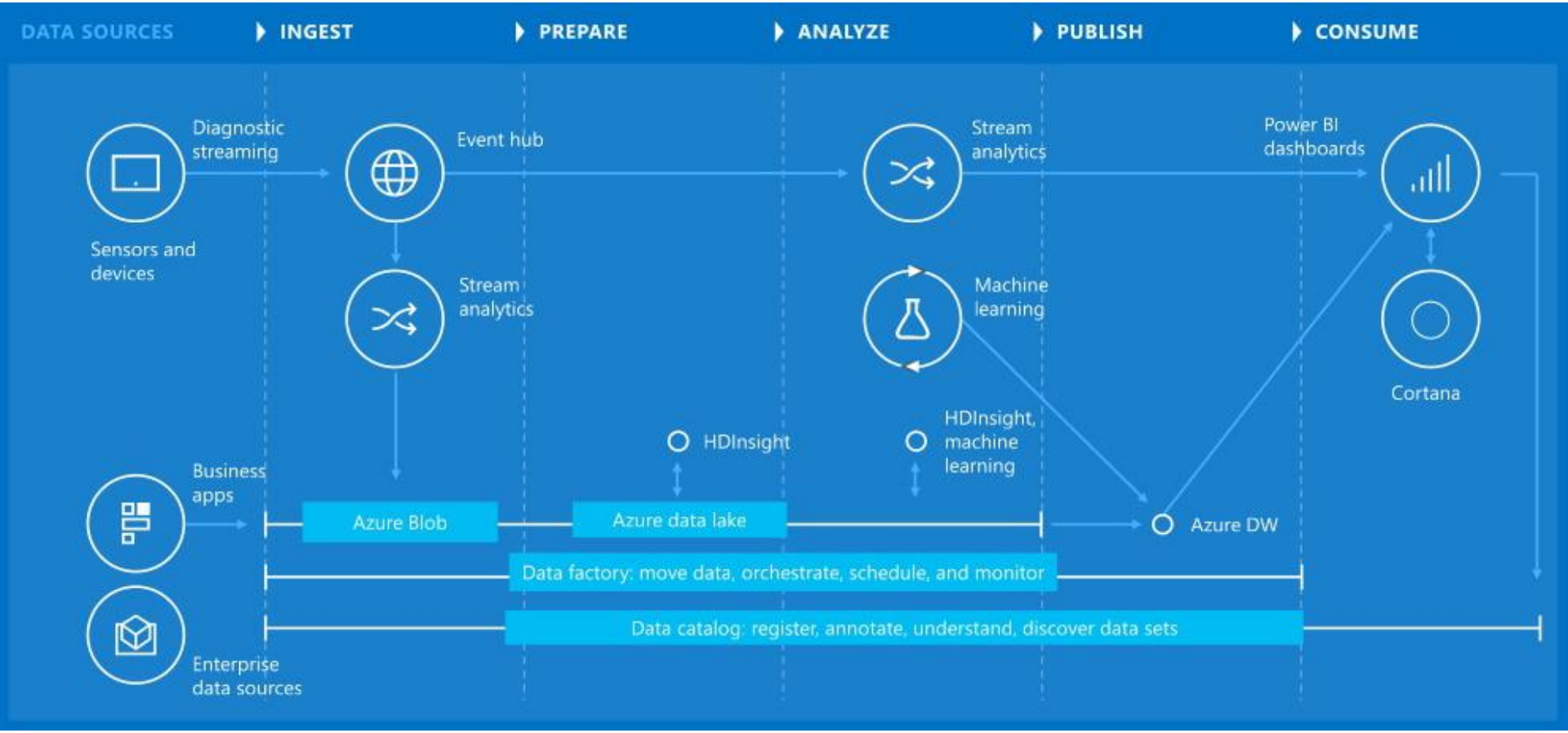
Machine Learning



Azure Machine Learning

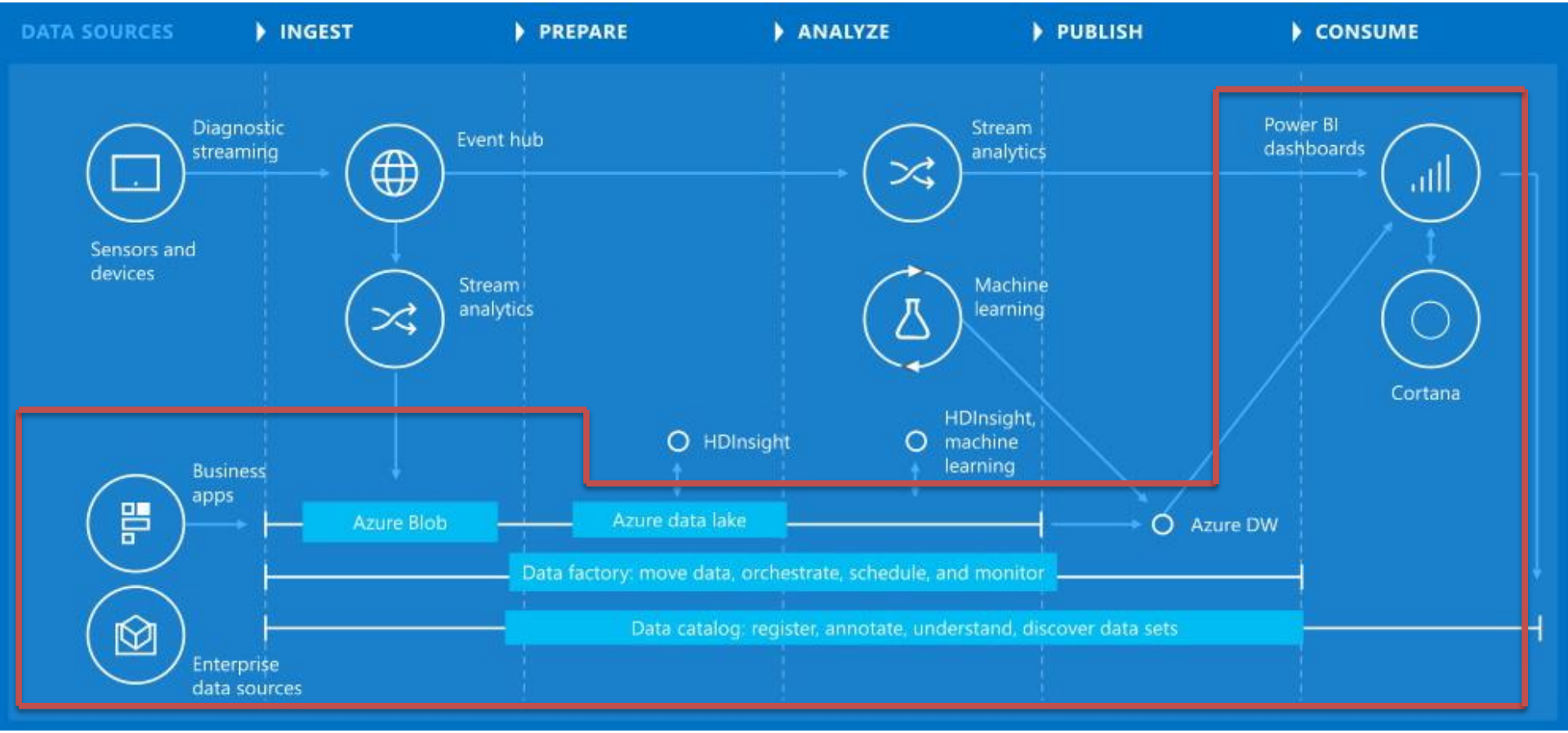


Cortana Intelligence



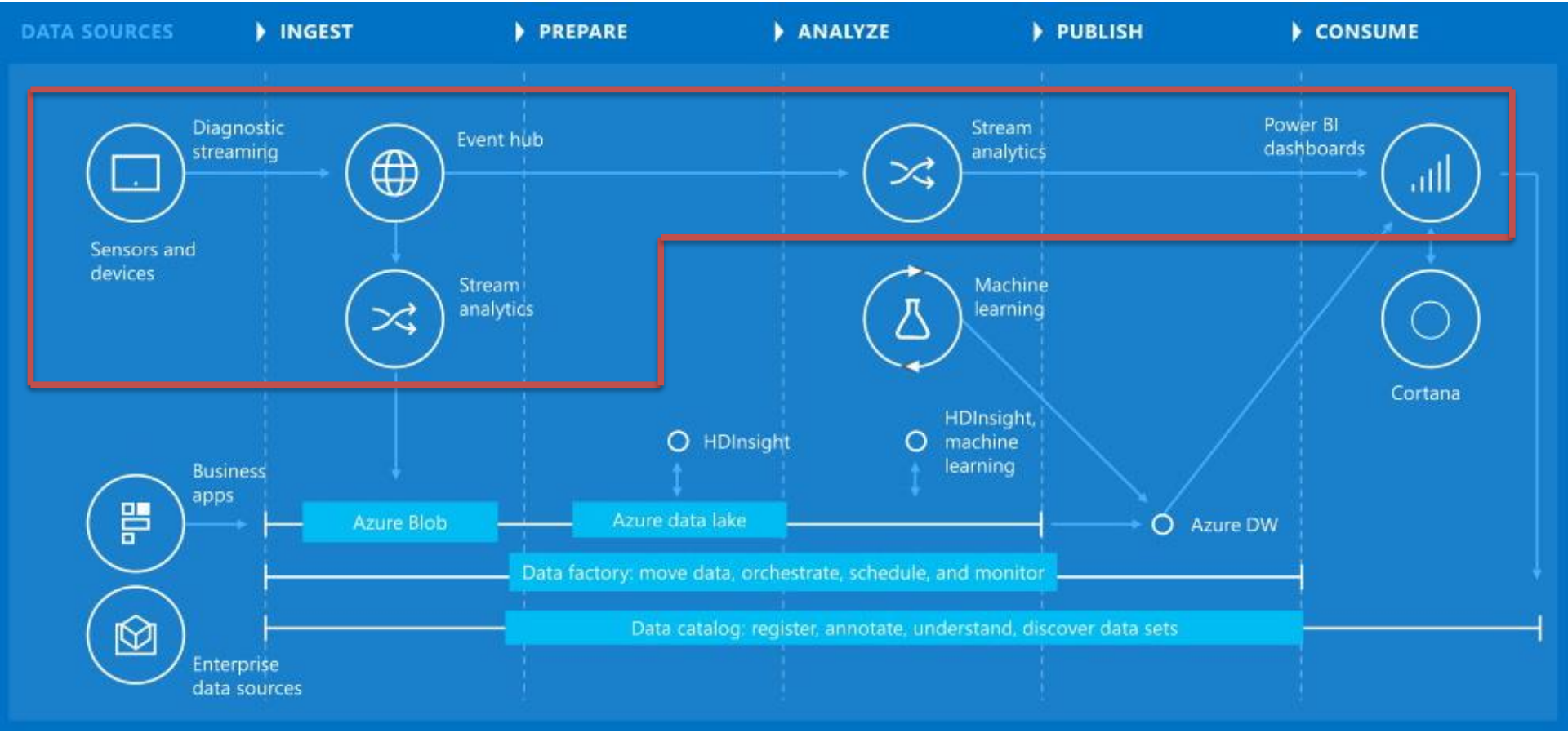
purplefrog
business intelligence consultancy

Cortana Intelligence - Batch



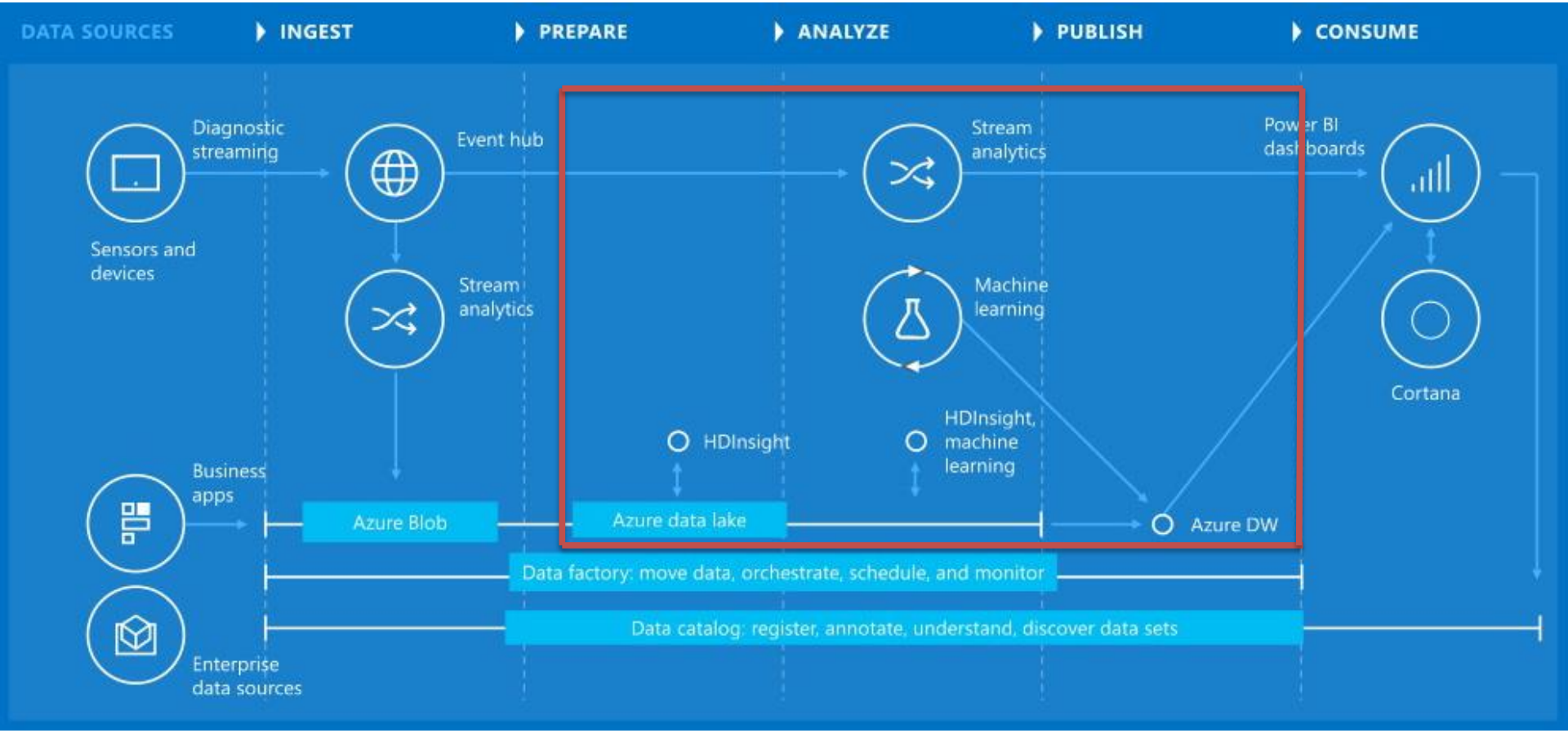
purplefrog
business intelligence consultancy

Cortana Intelligence - RealTime



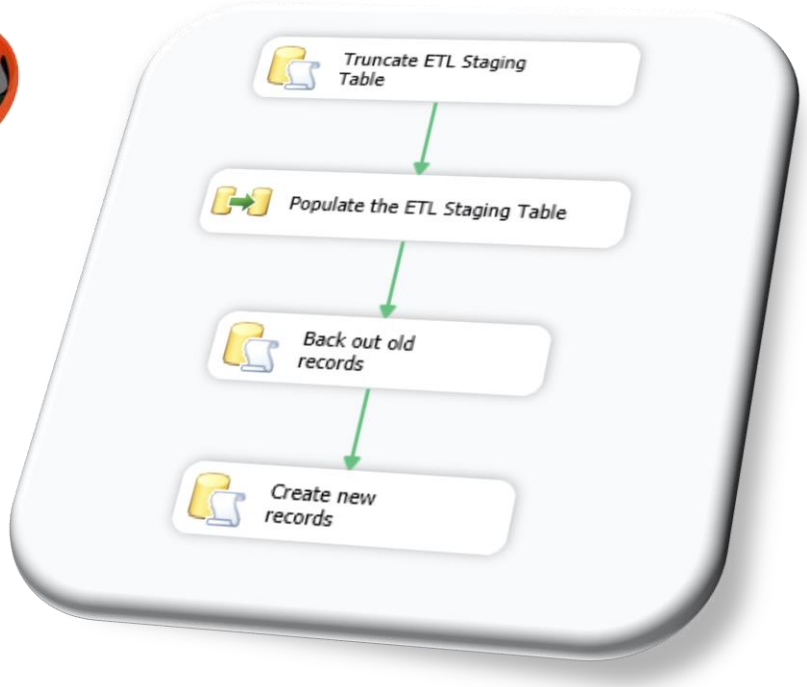
purplefrog
business intelligence consultancy

Cortana Intelligence – Analytics



purplefrog
business intelligence consultancy

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*

purplefrog
business intelligence consultancy

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);
```

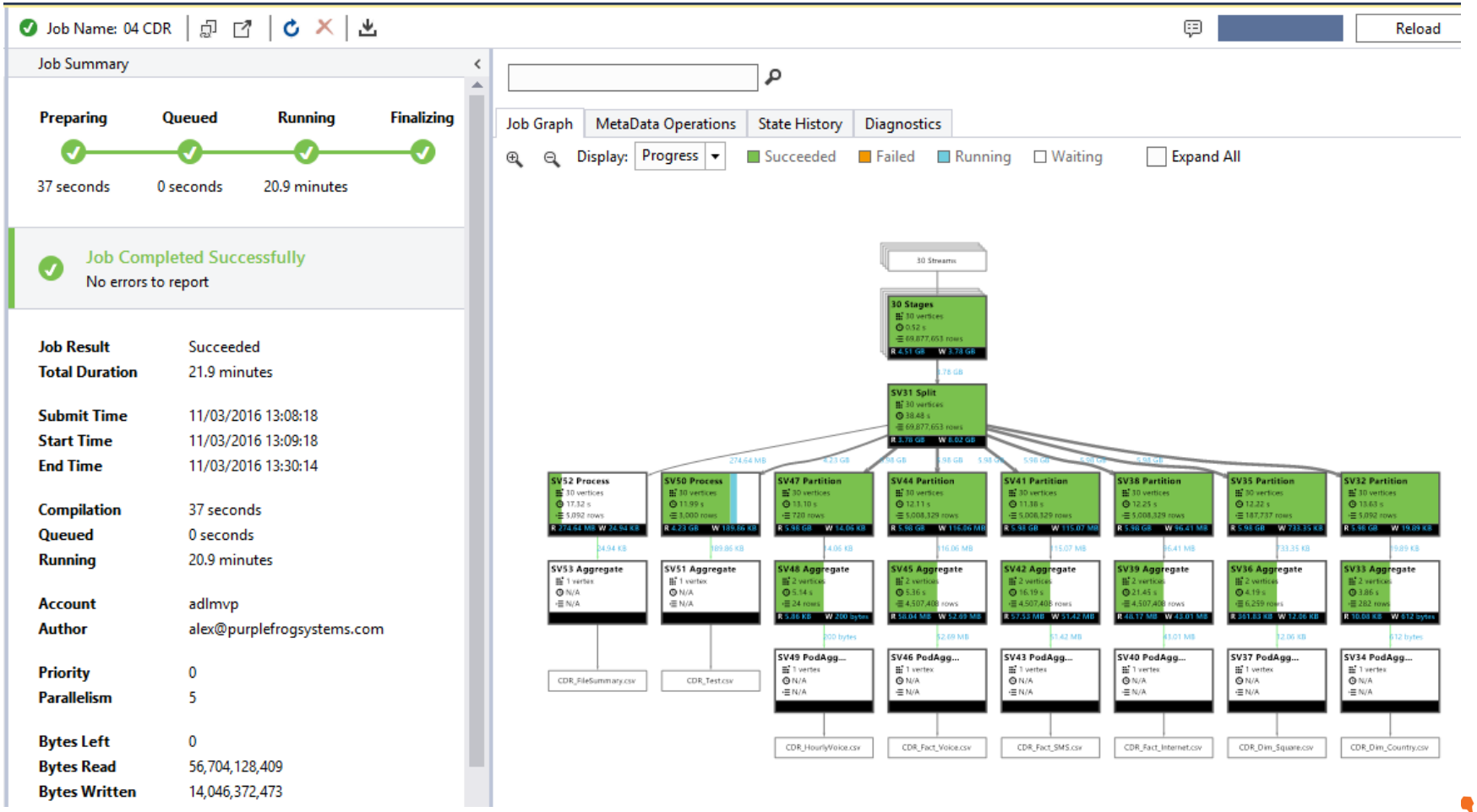
purplefrog
business intelligence consultancy

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 "";

}
```

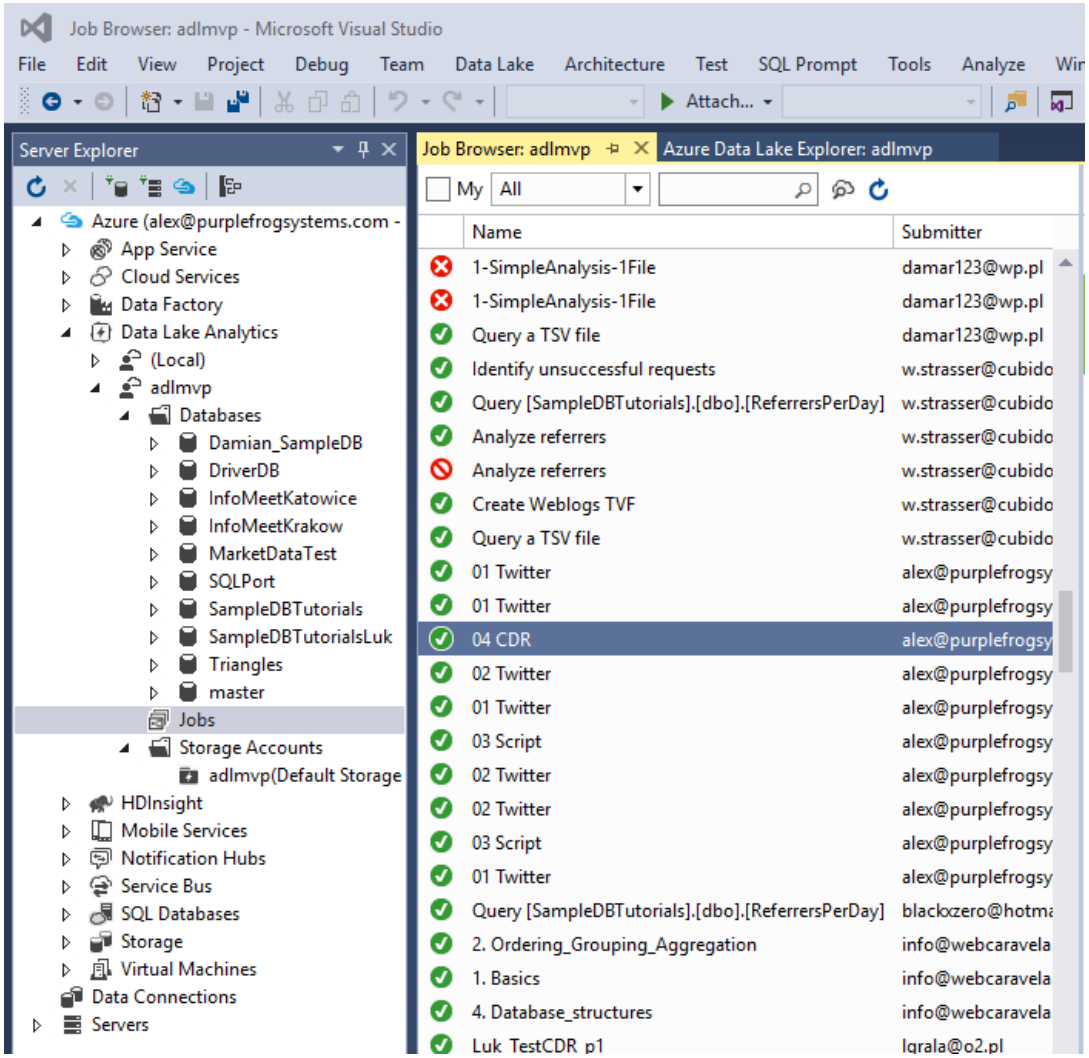
purplefrog
business intelligence consultancy

U-SQL / Data Lake



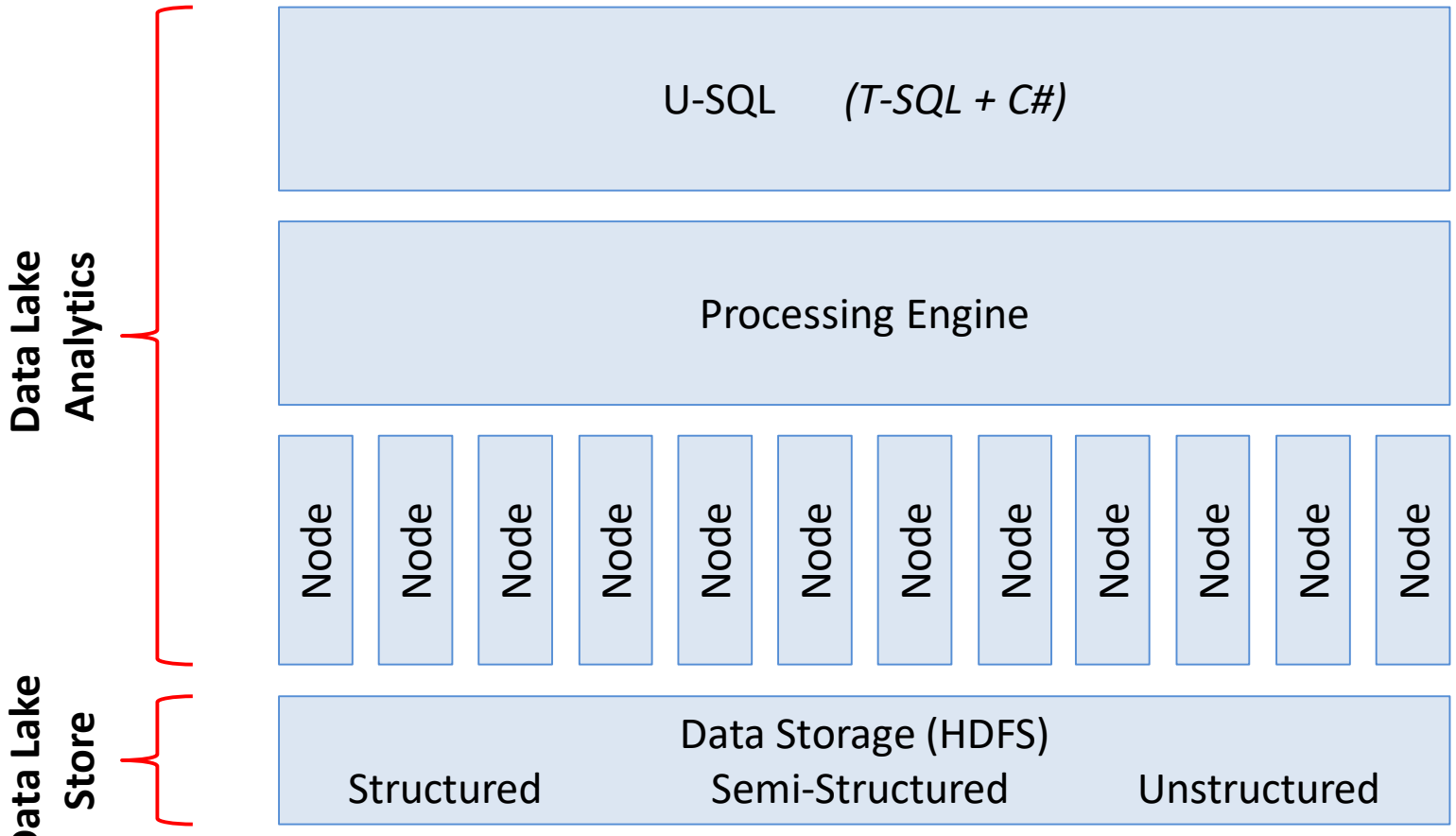
purplefrog
business intelligence consultancy

U-SQL / Data Lake



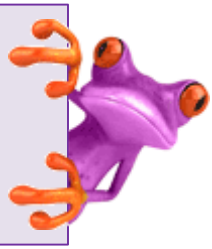
purplefrog
business intelligence consultancy

Data Lake

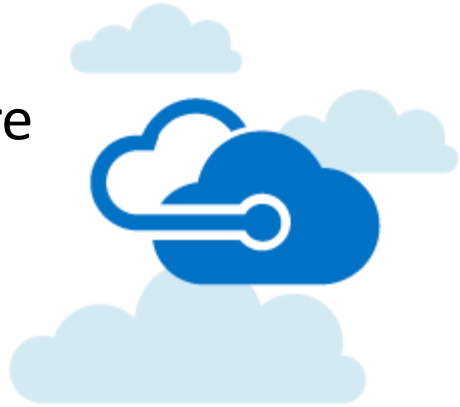


purplefrog
business intelligence consultancy

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



purplefrog
business intelligence consultancy

Business Intelligence in Azure

Alex@PurpleFrogSystems.com
@PurpleFrogSys
www.PurpleFrogSystems.com
www.PurpleFrogSystems.com/blog