AZURE DATA FACTORY – DATA FLOWS

BRINGING THE “T” IN ELT
WHO AM I?

• Practice Manager – Pragmatic Works – www.pragmaticworks.com
• Former CIO
• Blog, Speak, Record, Tweet
  • Twitter: @bizdataviz
  • LinkedIn: in/cseferlis
  • http://blog.bizdataviz.com
• MBA from UMass
• Ski, Bike, Hike, Run and drag the family along…

T: @bizdataviz
In/cseferlis
WHY ARE WE HERE?

• Modern Data Warehouse (Platform)
• Azure Data Orchestration
• Preview New Azure Data Factory Features
• Full Circle ELT with Azure Data Factory
AZURE DATA FACTORY… QUICKLY

- Native Cloud Data Orchestration Tool
- Started with v1, but lacked many features
- v2 goes GA 6/2018 – work has begun on Data Flows
- Data Flows and other new features currently in Limited Preview
LIMITED PREVIEW FEATURES

• Debug Mode
• Data Inspection
• Transformations
• GitHub Support
• Detailed Flow Inspection
• All-New Expression Builder
VISUAL DATA FLOW AUTHORING

- Transform Data, At Scale, in the Cloud, Zero-Code
  - Cloud-first, scale-out ELT
  - Code-free dataflow pipelines
- Serverless scale-out transformation execution engine
- Maximum Productivity for Data Engineers
  - Does NOT require understanding of Spark / Scala / Python / Java
- Resilient Data Transformation Flows
  - Built for big data scenarios with unstructured data requirements
  - Operationalize with Data Factory scheduling, control flow and monitoring
CODE-FREE DATA TRANSFORMATION AT SCALE

• Does not require understanding of Spark, Big Data Execution Engines, Clusters, Scala, Python ...

• Focus on building business logic and data transformation
  • Data cleansing
  • Aggregation
  • Data conversions
  • Data prep
  • Data exploration

... not ...

t: @bizdataviz
ln/cseferlis
ADF DATA FLOW WORKSTREAM

Data Sources
- Explicit user action
- User places data source(s) on design surface, from toolbox
- Select explicit sources

Staging
- Implicit/Explicit
- Data Lake staging area as default
- User does not need to configure this manually

Transformations
- Explicit user action
- User places transformations on design surface, from toolbox
- User must set properties for transformation steps

Destination
- Explicit user action
- User chooses destination connector(s)
- User sets connector property options

User places data source(s) on design surface, from toolbox
Select explicit sources

Sort, Merge, Join, Lookup …
• Azure SLAs are NA for preview services (private or public preview) until GA of the service.

• Limited Preview Support
  • Handled directly with the Azure Engineering team via adfdataflowext@microsoft.com.

• Sign-up for ADF Data Flow service
  • http://aka.ms/dataflowpreview
  • Microsoft Azure must whitelist your subscription ID to turn on the feature for you

• Public Preview Support
  • Normal Azure customer service channels

DATA FLOW LIMITED PREVIEW

t: @bizdataviz
ln/cseferlis
COMMON ELT SCENARIOS HANDLED WITH ADF – DATA FLOWS
SIMPLE COPY FLOW

- ADF Data Flow is a guided construction process
- Begin by defining the Datasets for your Source and Sink
- Add Transformations to each node in your data flow
- Or simply copy from source to sink with no transformation
- Map columns and fields along the way
SLOWLY CHANGING DIMENSION

- Common DW pattern to manage changing attributes to dimension members
- Graphically build code-free SCD ETL pattern to load your data warehouse
- Connect directly to Azure SQL DB and Azure SQL DW
- Use Lookup, Surrogate Key, Derived Column and Select transforms
LOAD STAR SCHEMA DW

- Classic ETL pattern is easy to build in ADF’s code-free Data Flow visual data transformation environment
- Add Aggregate transforms to produce calculations that you store in your analytical database schema
- Use Join transform to combine data from multiple data sources and data streams inside your data flow
- Land your data in your Lake folders or direct to Azure SQL DW
DATA LAKE/DATA SCIENCE

- ADF supports building visual data transformations against your data directly in Data Lake locations (i.e. Azure Blob Store, Azure Data Lake Store)
- Built-in handling of schema drift for frequent changes in data lake file formats, columns, and data types
- Perform data exploration and data profiling across your data lake in ADF Data Flow with interactive debug data preview
NEW FEATURES IN AZURE DATA FACTORY
INTERACTIVE EXPRESSION BUILDER – BUILD TRANSFORM EXPRESSIONS, NOT CODE

t: @bizdataviz
ln/cseferlis
DEBUG DATA FLOWS WITH
DATA PREVIEW AND DATA SAMPLING

Number of columns: 2
Number of rows: 30

Output schema: Data Preview

<table>
<thead>
<tr>
<th>Order</th>
<th>Date</th>
<th>Column</th>
<th>Profit</th>
<th>Continent</th>
<th>Column</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>12/03/2018</td>
<td>✔</td>
<td>2455.45</td>
<td>Cell Contents</td>
<td>North America</td>
</tr>
<tr>
<td>2</td>
<td>12/03/2018</td>
<td>✔</td>
<td>2455.45</td>
<td>Cell Contents</td>
<td>North America</td>
</tr>
<tr>
<td>3</td>
<td>12/03/2018</td>
<td>✔</td>
<td>2455.45</td>
<td>Cell Contents</td>
<td>North America</td>
</tr>
<tr>
<td>4</td>
<td>12/03/2018</td>
<td>✔</td>
<td>2455.45</td>
<td>Cell Contents</td>
<td>North America</td>
</tr>
<tr>
<td>5</td>
<td>12/03/2018</td>
<td>✔</td>
<td>2455.45</td>
<td>Cell Contents</td>
<td>Africa</td>
</tr>
<tr>
<td>6</td>
<td>12/03/2018</td>
<td>✔</td>
<td>2455.45</td>
<td>Cell Contents</td>
<td>Asia</td>
</tr>
<tr>
<td>7</td>
<td>12/03/2018</td>
<td>✔</td>
<td>null</td>
<td>Cell Contents</td>
<td>Europe</td>
</tr>
<tr>
<td>8</td>
<td>12/03/2018</td>
<td>✔</td>
<td>null</td>
<td>Cell Contents</td>
<td>North America</td>
</tr>
<tr>
<td>9</td>
<td>12/03/2018</td>
<td>✔</td>
<td>49582.23</td>
<td>Cell Contents</td>
<td>Asia</td>
</tr>
<tr>
<td>10</td>
<td>12/03/2018</td>
<td>✔</td>
<td>49582.23</td>
<td>Cell Contents</td>
<td>Asia</td>
</tr>
<tr>
<td>11</td>
<td>12/03/2018</td>
<td>✔</td>
<td>null</td>
<td>Cell Contents</td>
<td>Europe</td>
</tr>
<tr>
<td>12</td>
<td>12/03/2018</td>
<td>✔</td>
<td>49582.23</td>
<td>Cell Contents</td>
<td>Europe</td>
</tr>
<tr>
<td>13</td>
<td>12/03/2018</td>
<td>✔</td>
<td>49582.23</td>
<td>Cell Contents</td>
<td>North America</td>
</tr>
</tbody>
</table>

Number of columns: 12
Number of rows: 18

Continent_data

528 51.8% North America
203 20.3% Asia
136 13.6% Africa
116 11.6% Europe
17 1.7% Other

- Content Type: String
- Missing Value: 0 (0%)
- Unique Values: 5 (5%)
- Most Common: North America (52.8%)
- Skewness: -0.043210436
BUILD RESILIENT DATA FLOWS WITH SCHEMA DRIFT HANDLING
HANDLING SOURCE CHANGES

• Data Engineer Defines Source and you take ALL fields from source w/flexible schema
• Data Engineer derives columns using template expression patterns based on name and type matching. No need to define static field names.
Data Engineer derives columns using template expression based on name and type matching. No need to define static field names.
• Data Engineer derives columns using template expression based on name and type matching
Sink all incoming fields along with new derived field
IMPORTANT LINKS:

• Sign-up for ADF Data Flow service:
  • [http://aka.ms/dataflowpreview](http://aka.ms/dataflowpreview)
  • Microsoft Azure must whitelist your subscription ID to turn on the feature for you

• GitHub Repository for documentation:
  • [https://github.com/kromerm/adfdataflowdocs/](https://github.com/kromerm/adfdataflowdocs/)
THANK YOU!