



Acquiring Data Through ETL: Do it Right

Bala Mullachery

2019 ESRI DEVELOPER SUMMIT
Palm Springs, CA

Agenda

- What is ETL?
 - Why ETL?
 - What are the Best Practices?
 - How to ETL?
- 

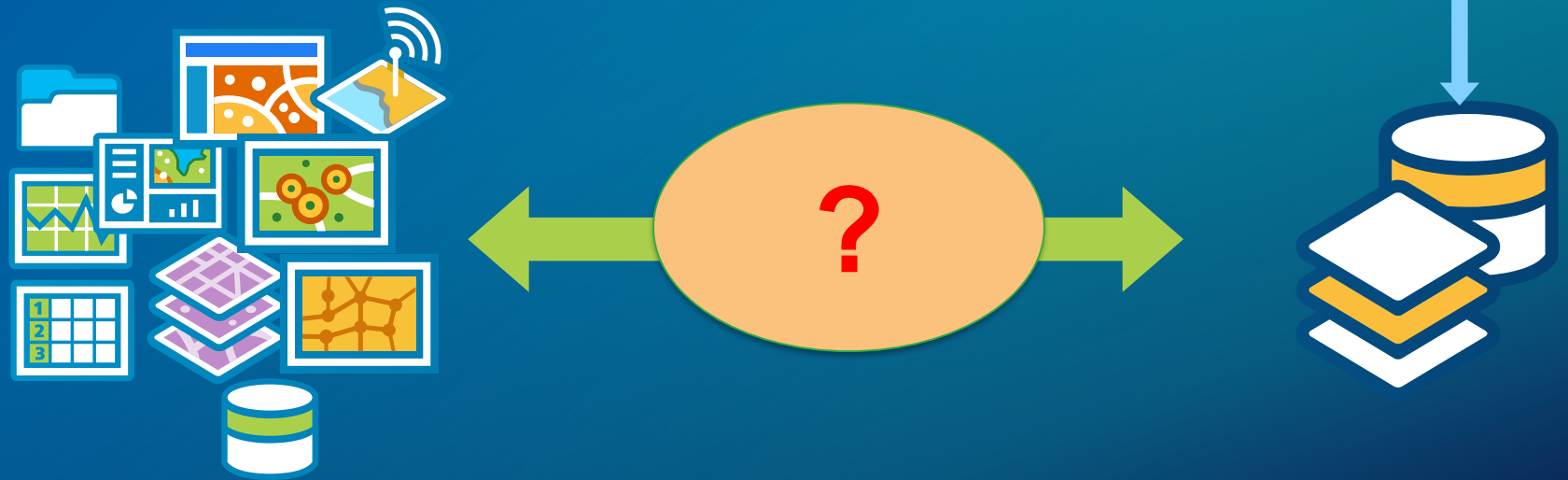
Data Format

- What formats?
- From A,B,C... to D,E,F... at once?
- Local data or web too?
- What can I do on the way?
- Is this part of Geoprocessing?
- Can I automate this on a schedule?
- Do I have to code?



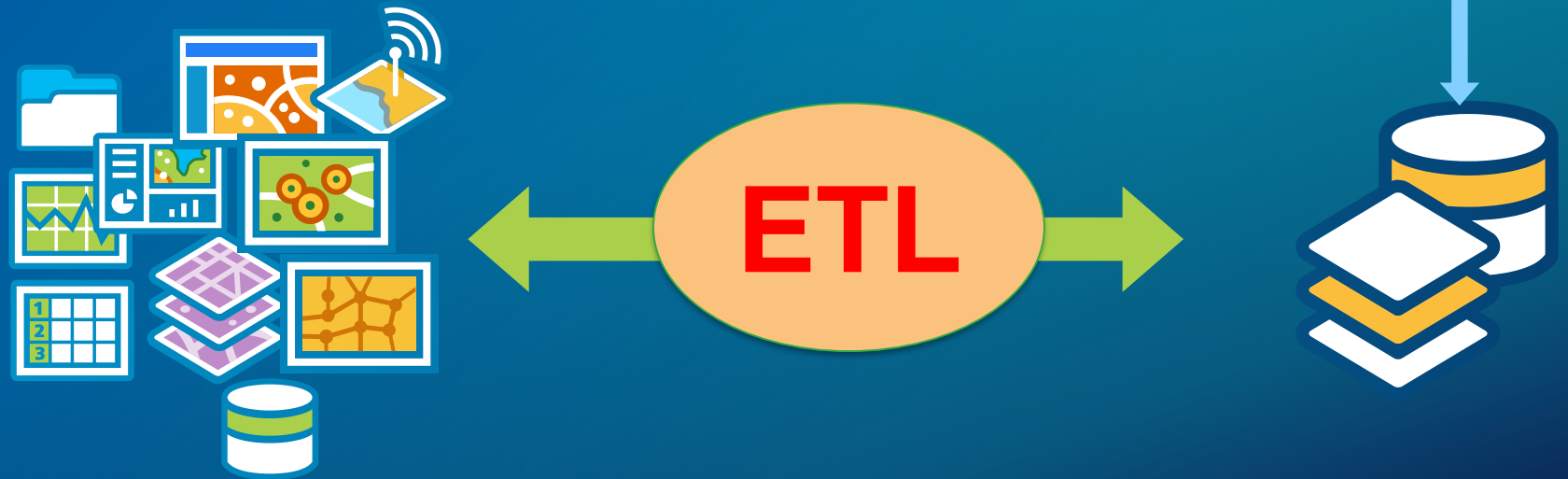
What is ETL?

- **E**xtract > **T**ransform > **L**oad
- **Extract** – connect to data source(s) and withdraw data
- **Transform** – manipulate data
- **Load** – load data to the destination



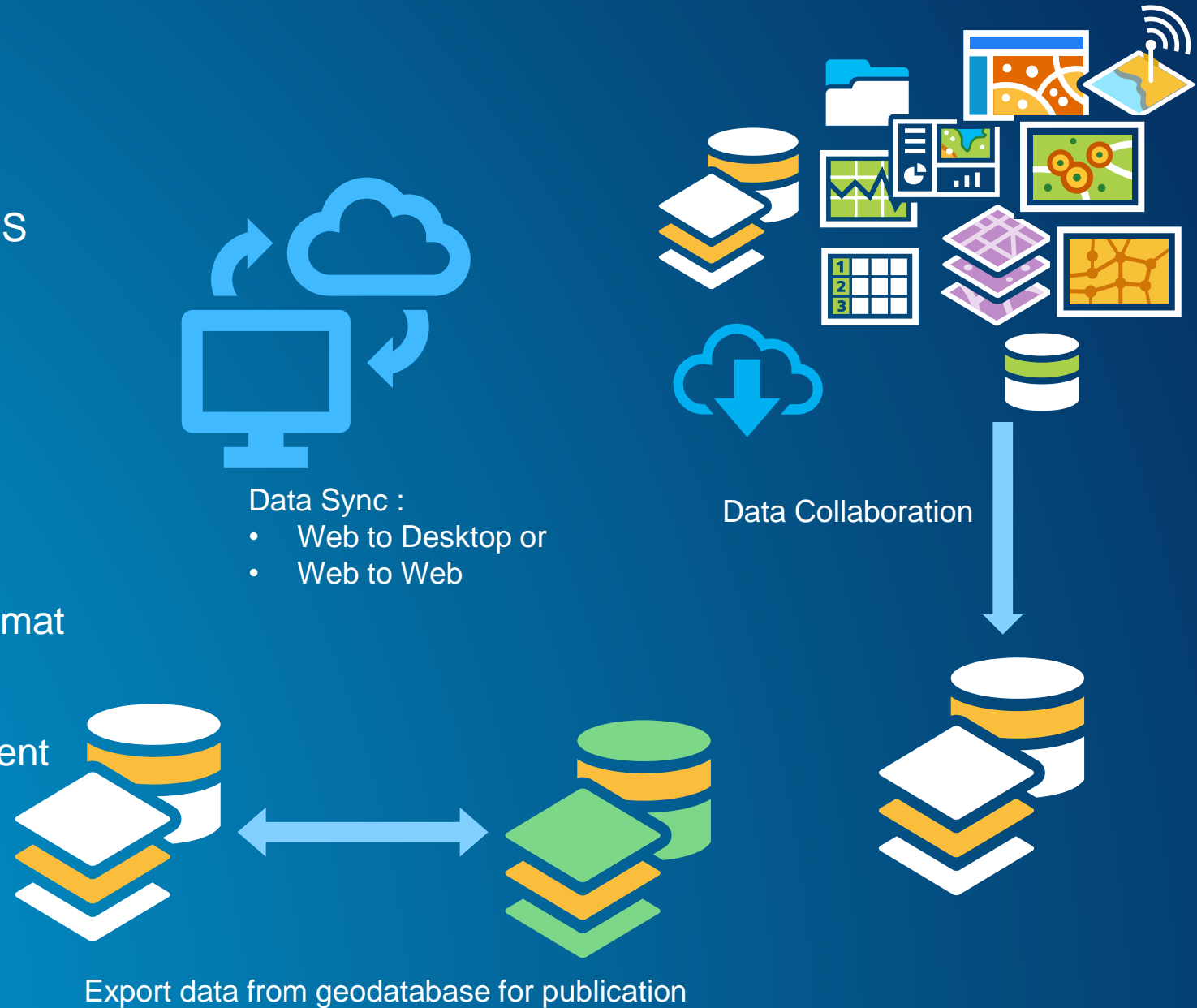
What is ETL? ...

- **E**xtract > **T**ransform > **L**oad
- **Extract** – connect to data source(s) and withdraw data
- **Transform** – manipulate data
- **Load** – load data to the destination



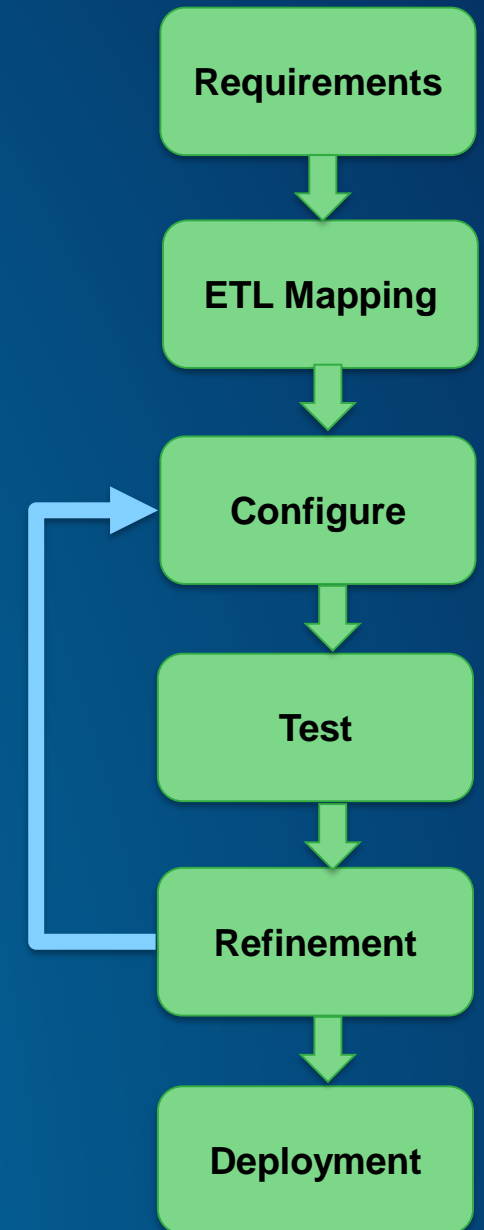
Why ETL? Use Cases

- Migration
 - Adoption of ArcGIS, Working with ArcGIS
- Standardization
 - Data cleanup for advanced analytics
- Connection
 - Apps with an API
- Collaboration
 - Generate data in a business-friendly format
- Validation
 - Feature-Level Business Rule Enforcement
- Detection
 - Cross-Format Feature Interaction
- Synchronization
 - Feature Services and other repositories



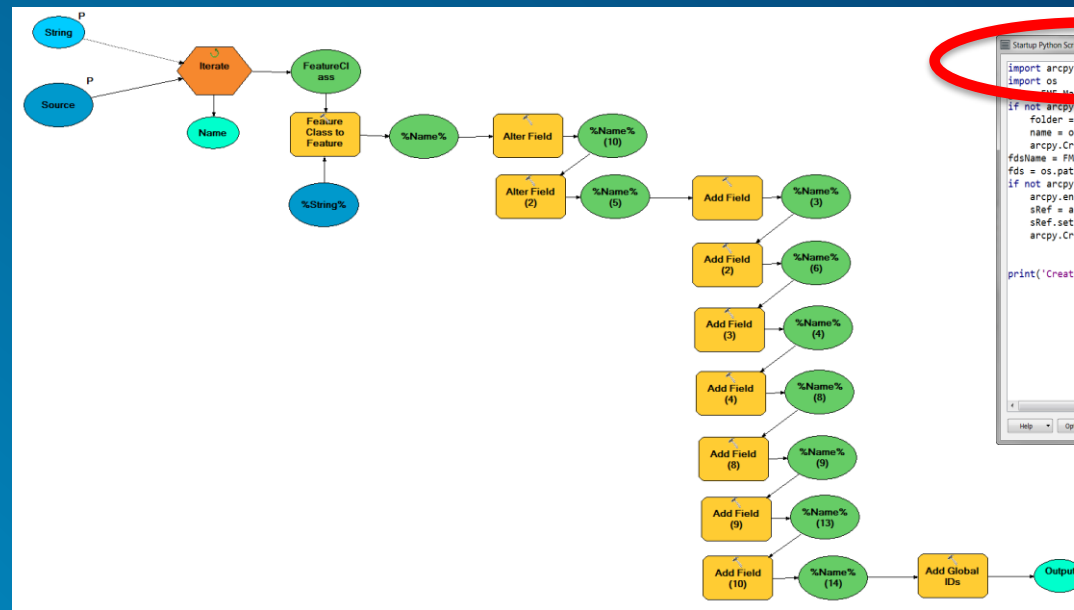
How ETL? Recommended Approach

- **Gather requirements**
 - Source(s) data format
 - Destination data format
 - Destination data requirements
 - Source data anomalies
 - Feature editing and data enhancement
 - New data creation
- **Prepare ETL mapping document**
 - Source to Destination feature mapping
 - Feature data rules
- **Configuration and Testing**
 - Technology to use
- **Deployment**
 - On-demand
 - Batch processing / Scheduler



How ETL? Technology

- ArcCatalog, ArcMap
- ArcGIS Model builder
 - Geo-processing tools
- Python
- Interoperability



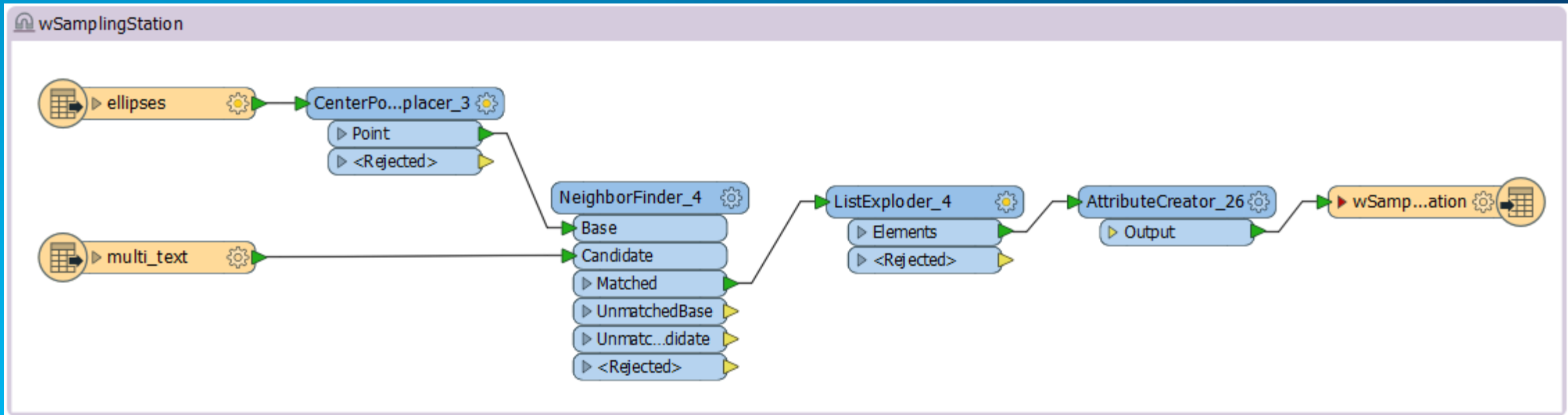
```
import arcpy
import os

gdb = arcpy.GetMacroValue('GDB_PATH')
DATABASE_FILE = 'DATABASE_FILE'

if not arcpy.Exists(gdb):
    folder = os.path.dirname(gdb)
    name = os.path.basename(gdb)
    arcpy.CreateFileGDB_management(out_folder_path=folder, out_name=name, out_version='CURRENT')
    fdsName = arcpy.GetMacroValue('GEOID_FEATURE_DATASET')
    fds = os.path.join(gdb, fdsName)
    if not arcpy.Exists(fds):
        arcpy.env.zResolution = '0.0001 meters'
        sRef = arcpy.SpatialReference(4326)
        sRef.setDomain(-6378137, 10000)
        arcpy.CreateFeatureDataset_management(out_dataset_path=gdb,
            out_name=fdsName,
            spatial_references=sRef)
    print('Created Feature Dataset: {}'.format(fdsName))
```

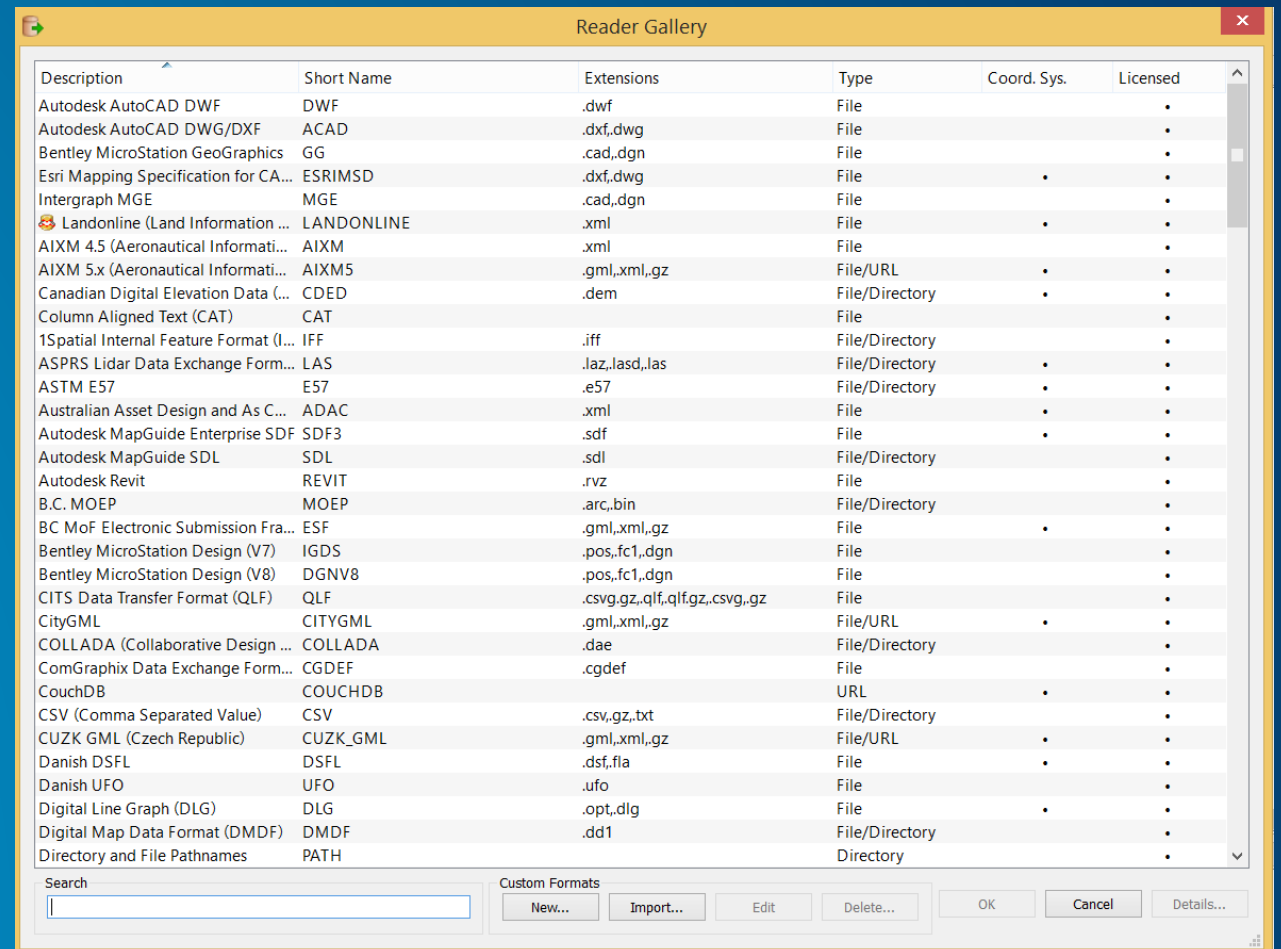
```
# coding: utf-8
tbxPath = r'C:\Work\UC2018\Presentations\Data Interoperability Extension TW\DataInteroperabilityTW.tbx'
arcpy.ImportToolbox(tbxPath)
quakeURL = r'https://earthquake.usgs.gov/earthquakes/feed/v1.0/summary/all_day.geojson'
kmlPath = r'C:\Work\UC2018\Presentations\Data Interoperability Extension TW\AllDayQuakes.kml'
result = arcpy.AllDayQuakes_DataInteroperabilityTW(arcpy.ImportToolbox(tbxPath))
print('\nCreated KML file: {}'.format(result.getOutput(0)))

Created KML file: C:\Work\UC2018\Presentations\Data Interoperability Extension TW\AllDayQuakes.kml
```



Data Interoperability – Source Format

- Read source data existing in various format:
 - CAD
 - GIS/Spatial data
 - Tabular/Spreadsheet
 - Text/CSV



The image shows a screenshot of the 'Reader Gallery' window in a software application. The window contains a table with columns for Description, Short Name, Extensions, Type, Coord. Sys., and Licensed. The table lists various data formats such as AutoCAD DWF, Bentley MicroStation GeoGraphics, Esri Mapping Specification for CAD, Intergraph MGE, Landonline (Land Information System), AIXM 4.5, AIXM 5x, Canadian Digital Elevation Data, Column Aligned Text (CAT), 1Spatial Internal Feature Format, ASPRS Lidar Data Exchange Format, ASTM E57, Australian Asset Design and As C..., Autodesk MapGuide Enterprise SDF, Autodesk MapGuide SDL, Autodesk Revit, B.C. MOEP, BC MoF Electronic Submission Format, Bentley MicroStation Design (V7), Bentley MicroStation Design (V8), CITS Data Transfer Format (QLF), CityGML, COLLADA (Collaborative Design), ComGraphix Data Exchange Format, CouchDB, CSV (Comma Separated Value), CUZK GML (Czech Republic), Danish DSFL, Danish UFO, Digital Line Graph (DLG), Digital Map Data Format (DMDF), and Directory and File Pathnames. At the bottom of the window, there is a search bar and a 'Custom Formats' section with buttons for 'New...', 'Import...', 'Edit', 'Delete...', 'OK', 'Cancel', and 'Details...'.

Description	Short Name	Extensions	Type	Coord. Sys.	Licensed
Autodesk AutoCAD DWF	DWF	.dwf	File		•
Autodesk AutoCAD DWG/DXF	ACAD	.dxf,dwg	File		•
Bentley MicroStation GeoGraphics	GG	.cad,dgn	File		•
Esri Mapping Specification for CA...	ESRIMSD	.dxf,dwg	File	•	•
Intergraph MGE	MGE	.cad,dgn	File		•
Landonline (Land Information ...	LONDONLINE	.xml	File	•	•
AIXM 4.5 (Aeronautical Informati...	AIXM	.xml	File		•
AIXM 5x (Aeronautical Informati...	AIXM5	.gml,xml,gz	File/URL	•	•
Canadian Digital Elevation Data (...	CDED	.dem	File/Directory	•	•
Column Aligned Text (CAT)	CAT		File		•
1Spatial Internal Feature Format (I...	IFF	.iff	File/Directory		•
ASPRS Lidar Data Exchange Form...	LAS	.laz,.lasd,.las	File/Directory	•	•
ASTM E57	E57	.e57	File/Directory	•	•
Australian Asset Design and As C...	ADAC	.xml	File	•	•
Autodesk MapGuide Enterprise SDF	SDF3	.sdf	File	•	•
Autodesk MapGuide SDL	SDL	.sdl	File/Directory		•
Autodesk Revit	REVIT	.rvz	File		•
B.C. MOEP	MOEP	.arc,.bin	File/Directory		•
BC MoF Electronic Submission Fra...	ESF	.gml,xml,gz	File	•	•
Bentley MicroStation Design (V7)	IGDS	.pos,.fc1,.dgn	File		•
Bentley MicroStation Design (V8)	DGNV8	.pos,.fc1,.dgn	File		•
CITS Data Transfer Format (QLF)	QLF	.csv,gz,.qlf,.qlf.gz,.csvg,.gz	File		•
CityGML	CITYGML	.gml,xml,gz	File/URL	•	•
COLLADA (Collaborative Design ...	COLLADA	.dae	File/Directory		•
ComGraphix Data Exchange Form...	CGDEF	.cgdef	File		•
CouchDB	COUCHDB		URL	•	•
CSV (Comma Separated Value)	CSV	.csv,gz,.txt	File/Directory		•
CUZK GML (Czech Republic)	CUZK_GML	.gml,xml,gz	File/URL	•	•
Danish DSFL	DSFL	.dsf,.fla	File	•	•
Danish UFO	UFO	.ufo	File		•
Digital Line Graph (DLG)	DLG	.opt,.dlg	File	•	•
Digital Map Data Format (DMDF)	DMDF	.dd1	File/Directory		•
Directory and File Pathnames	PATH		Directory		•

Data Interoperability - Transformation

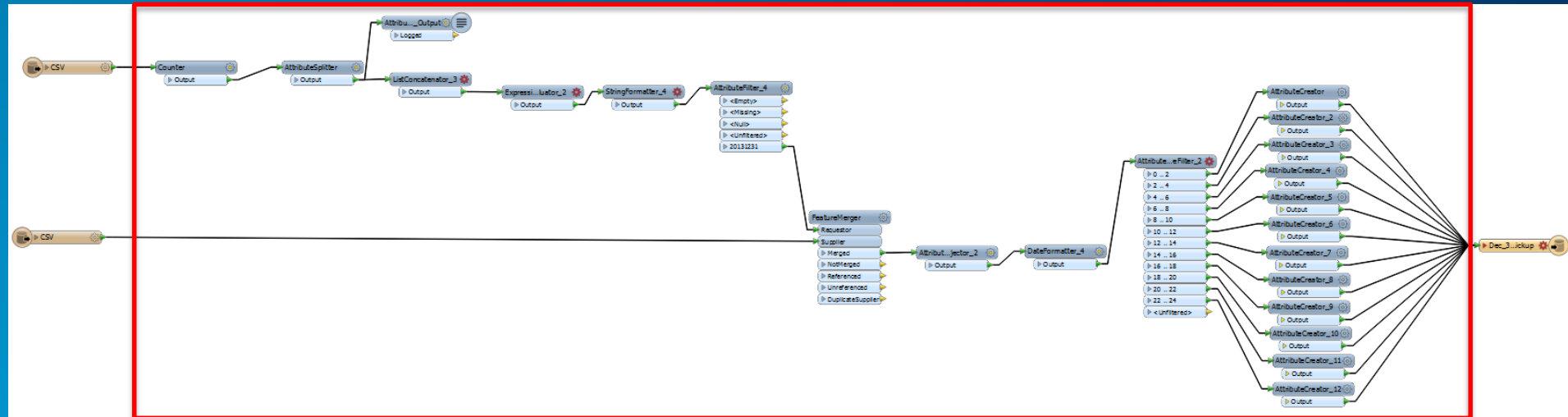
- Manipulate/Create data

- Spatial data

- Create new features
- Topology cleanup
- Feature matching
- Feature merger
- Projection

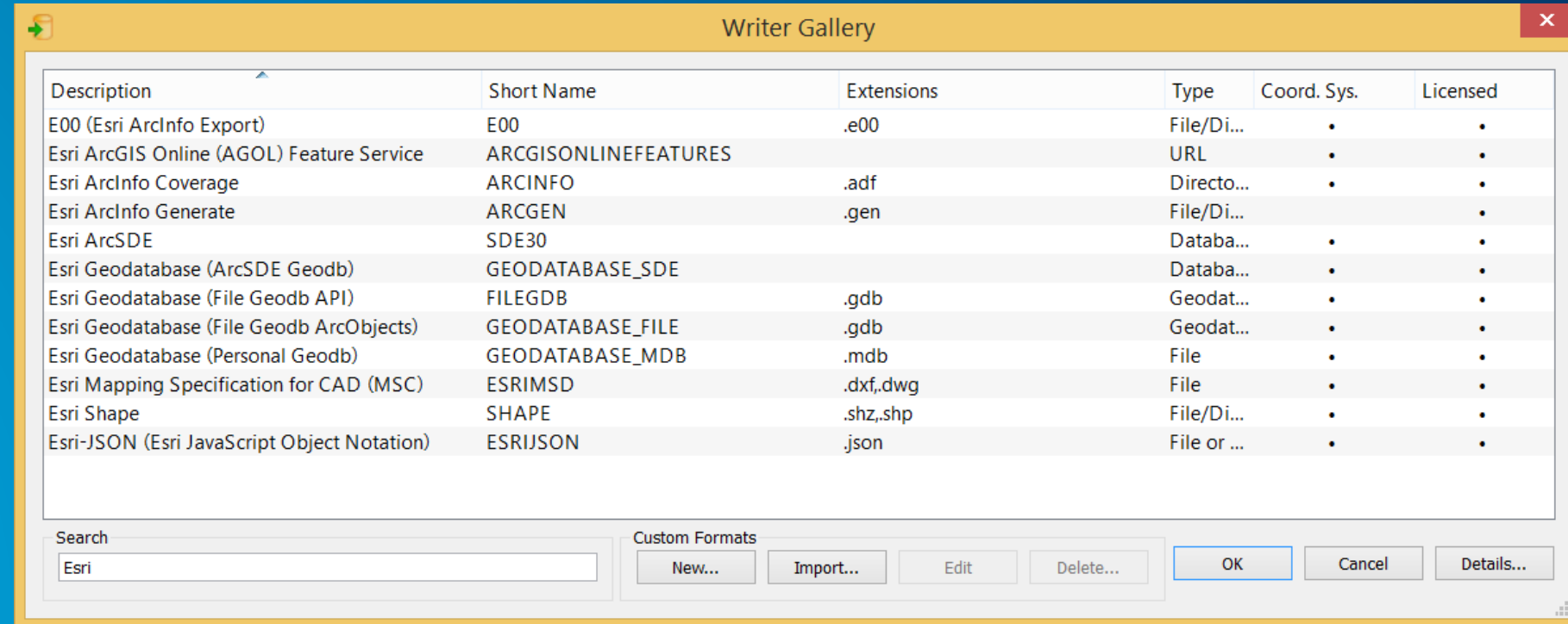
- Attribute data

- Domain values
- Unique identifier
- Tables join
- External Database Integration
- SQL Query



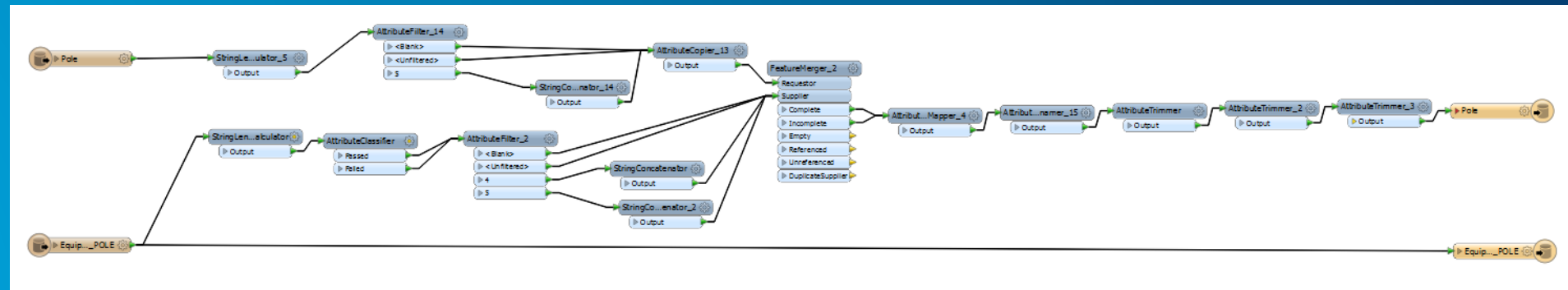
Data Interoperability – Destination Format

- Load data into geodatabase
 - Format conversion
 - Transactional or non-transactional data
 - Feature classes
 - Tables
 - Annotations
 - Geometric networks



Data Interoperability – Test and Refinement

- Feature count
- Visual validation
- Automated validation using Data Reviewer
- Refinement
- Release for production
 - On-demand
 - Scheduled intervals



Best Practices (recap)

- **Gather requirements**
- **Develop ETL mapping**
 - Source to destination
 - Rules
 - Source data issues
- **Configure ETL**
- **Test and Refinement**
- **Deployment Scenarios**
- **Visit:**
 - <https://community.esri.com/videos/5108-arcgis-data-interoperability-in-action>

• Question ?



esri

THE
SCIENCE
OF
WHERE