Stuart Tomlinson
Microsoft Cloud Consultant

Discover Big Data with Azure Data Factory

CompanyNet

company-net.com

CompanyNet

company-net.com

Agenda

What is Azure Data Factory

Data Integration

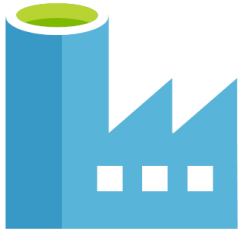
Data Factory Components

Demonstration

Big Data in the Cloud:

What is Data Factory?

CompanyNet



What is **Azure Data Factory**?

2016- Introduced into Azure Preview (V1)
And was initially marketed as a serverless SSIS platform in the Azure cloud.

2020- Microsoft introduced ADF V2 with many more capabilities, with code free dataflows with over 90 connectors for data.

Azure Data Factory now Natively Integrates SQL, HD Insights, Oracle, Data Bricks and Synapse Big Data tools and many more, while still allowing SSIS packages to be migrated into ADF.

WHAT IS AZURE DATA FACTORY

READ THE DOCS

A CLOUD-BASED DATA INTEGRATION SERVICE THAT ORCHESTRATES DATA MOVEMENT & TRANSFORMATION BETWEEN DIVERSE DATA SOURCES ⟷ CLOUD COMPUTE RESOURCES AT SCALE

DATA INTEGRATION CHALLENGE

WHAT PROBLEM DOES IT SOLVE?

BIG DATA — DIVERSE DATA SOURCES, DURATIONS — STORED DATA CLEANED & ENRICHED — ACTIONABLE INSIGHTS

DATABASES, FILES, DEVICES, DATA STREAMS → DATA INGEST → DATA TRANSFORMATION → DASHBOARDS, REPORTS

DATA OPS TEAM — HELP! We need to collect all this data! — DATA MOVEMENT → DATA ANALYSTS — WE NEED DATA WITH RELEVANT CONTEXT TO DO OUR ANALYSIS

HOW CAN WE ORCHESTRATE (DATA MOVEMENT) AND OPERATIONALIZE (WORKFLOW)

AZURE DATA FACTORY

- Raw, unorganized data stored in relational, non-relational, and other storage systems
- We need actionable business insights but raw data lacks the correlating context (by itself) for analysis
- Big data needs a scalable service than can ORCHESTRATE the data movement and OPERATIONALIZE the data processing workflows!

THE SCENARIO YOU RECOGNIZE

GAME DEV COMPANY

- HAVE: Petabytes of game log data in the cloud
- WANT: Actionable insight into my customers
  - gaming preferences
  - player demographics
  - user behaviors
- USES
  - develop new features
  - improve player experience
  - upsell, cross-sell and drive business growth

DATA ANALYST

WHAT I NEED...

- * reference data from our on-prem services
  - customer info
  - marketing campaign
  - game metadata
- * combined with gameplay data logs from cloud
  - game play behaviors
  - success, failure rates etc.

WHAT I MUST DO

- * collect data from sources
- * process data to transform it
- * publish data into warehouse for analysts

AUTOMATE THIS WORKFLOW (TRIGGER IT BY AN EVENT OR MANUALLY — MONITOR & MANAGE DAILY SCHEDULES)

HOW DOES ADF WORK?

CAN YOU TELL ME WHAT ADF DOES AT A HIGH LEVEL?

- AZURE DATA FACTORY IS A SET OF INTERCONNECTED SYSTEMS.
- THAT GIVE YOU AN END-TO-END PLATFORM...
- FOR YOUR DATA ENGINEERING!

LET'S BREAK IT DOWN

DESIGN WIZARD (OR) CODE + COMPUTE

CONNECT & COLLECT

1. Build an information production system to connect all data sources and adapt ingest to their diverse intervals and speeds! (CONNECTORS)
2. Collect all data in a centralized location to facilitate processing (e.g. transformation) next (COPY ACTIVITY)

TRANSFORM & ENRICH

3. Transform the collected data using actions to aggregate, filter, clean etc. Use code-free UI based mapping data transformation graph creation (OR) use compute to transform data by hand. (MAPPING DATA FLOW)

CI/CD & PUBLISH

4. Manage data pipeline ops using Azure DevOps and GitHub! Incrementally develop and deploy your E-T-L process then publish processed data into Azure Data Warehouse, Azure SQL Database, Azure Cosmos DB – or your BI analytics engine! (PIPELINE OPS)
5. Once your data pipeline is created and deployed, monitor it for the success/failure rates! (AZURE MONITOR)

MONITOR & ALERT

Use Azure Monitor API, PowerShell, health panels on Azure Portal!

LET'S TALK DATA INTEGRATION PATTERNS

The most common pattern is ETL

- EXTRACT = connect to sources, copy data to central store
- TRANSFORM = process the data for analysis, data transformation at scale
- LOAD = move the data to data warehouse or analytics engines for business insights

Another pattern is ELT where raw (native) data is itself loaded (stored) before the transformation phase....

AZURE DATA FACTORY HAS CODE-FREE ETL AS A SERVICE!

You always have the option to do hand-coded transformation using Azure compute... but mapping data flows provide a UI-based wizard to simplify your pipeline setup..

5 ASPECTS IT COVERS!

1. INGEST DATA
   - * Copy data on-prem or in cloud!
   - * 90+ native connectors
   - * Serverless, auto-scaling!
2. CONTROL FLOW
   - * Design (via UI) or Create (via SDK) your data pipelines
   - * Utilize workflow constructs (loops, parameters, variables...) for efficiency
3. DATA FLOW
   - * Code-free data transforms (execute automagically on Spark)
   - * Scale out with Azure Integration Runtimes, generate data flows with SDK (code) or Designer (UI)
4. SCHEDULE OPS
   - * Build/Maintain ops schedules for your data pipelines
   - * Options include: Wall clock, Event based, Tumbling windows, chaining
5. MONITOR OPS
   - * View active executions and pipeline history – details at activity/pipeline
   - * Establish alerts for key events or progress notifications

7 THINGS TO KNOW ABOUT AZURE DATA FACTORY

AZURE DATA FACTORY

1. ENTERPRISE READY — Data Integration At cloud scale!
2. ENTERPRISE DATA READY — 90+ connectors! It just works.
3. CODE-FREE TRANSFORMATION — UI driven mapping data flows
4. RUN CODE ON ANY AZURE COMPUTE — For hands-on data transformations
5. MANY SSIS PACKAGES RUN ON AZURE — (Rehost on-prem in ADF in 3 steps)
6. ADF CAN MAKE DATA OPS SEAMLESS — source control, automated deploy, simple templates
7. SECURE DATA INTEGRATION — Managed virtual networks protect against data exfiltration, simplify your networking!

6 INTEGRATION RUNTIME (IR)

INTEGRATED RUNTIME (IR) IS COMPUTE INFRASTRUCTURE USED BY ADF TO PROVIDE FULLY MANAGED

- DATA FLOWS (transformation)
- DATA MOVEMENT (movement)
- ACTIVITY DISPATCH (route to service)
- SSIS PACKAGE EXECUTION (in managed compute)

3 IR TYPES

IR Types		CAPABILITY	N/W
AZURE-SSIS	* SSIS package execution (public or private)		
SELF HOSTED	* Activity Dispatch * Data Movement		public or private
AZURE	* Activity Dispatch * Data Movement * Data Flow		public or private

- SELF-HOSTED
- AZURE-SSIS
- AZURE

POTENTIAL TO HAVE HYBRID DATA INTEGRATION SOLUTIONS

PICK WHAT WORKS FOR YOUR DATA NEEDS ... AND YOUR NETWORKING REQUIREMENTS!

2 ACTIVITIES

- REPRESENT A SINGLE PROCESSING STEP IN THE PIPELINE e.g.
  - COPY ACTIVITY — copy data from one store (source) to other (destination)
  - HIVE ACTIVITY — run a Hive query on an HDInsight cluster
- THREE TYPES OF ACTIVITY SUPPORTED BY ADF
  - Data Movement
  - Data Transformation
  - Control

1 PIPELINES

A1 A2 A3 — PIPELINE STAGE

- A PIPELINE IS A LOGICAL GROUPING OF ACTIVITIES THAT PERFORM ONE UNIT OF WORK
- AN ADF INSTANCE CAN HAVE ONE OR MORE PIPELINES!

Example: Pipeline that ingests data from Azure Blob (A1), runs Hive query on HDInsight to partition it (A2) and moves result to next store (A3)

Benefit: Manage correlated activities as a single entity!

Execution: Activities may be chained (run sequentially) or be independent (run in parallel) within pipeline!

3 DATASETS

- REPRESENTS DATA STRUCTURE GIVING SELECT VIEW INTO DATA STORE
- POINTS TO THE SPECIFIC DATA SUBSET TO USE IN ACTIVITY...
- .. FOR INPUT & OUTPUT

DATA STORE — DATASET (VIEW)

4 LINKED SERVICES

- TWO USES
  - REPRESENT DATA SOURCE
  - REPRESENT COMPUTE RESOURCE
- REPRESENT CONNECTION INFORMATION TO LINK ADF TO EXTERNAL SERVICES
- WORKS CLOSELY WITH DATASETS. THIS SPECIFIES HOW TO CONNECT & DATASET DEFINES WHAT DATA LOOKS LIKE...

data sources for INGEST; compute resources for TRANSFORM

5 MAPPING DATA FLOWS

ADF EXECUTES LOGIC ON SPARK CLUSTER FOR YOU (automagically) NO NEED TO MANAGE OR MAINTAIN YOUR OWN CLUSTER!

- CREATE & MANAGE DATA TRANSFORMATION GRAPHS THAT WORK ON ANY SIZE DATA
- BUILD UP REUSABLE LIBRARY OF DATA TRANSFORMATION ROUTINES!
- EXECUTE TRANSFORM ROUTINES IN PIPELINES (AS ACTIVITY) TO SCALE!

HEY PRESTO!

5 ADDITIONAL CONCEPTS

1 CONTROL FLOW

REFERS TO THE ORCHESTRATION OF PIPELINE ACTIVITIES INCLUDING:

- CHAINING ACTIVITIES
- DEFINING PARAMETERS
- BRANCHING ACTIVITIES
- PASSING ARGUMENTS (for pipeline run)
- CUSTOM STATE PASSING
- LOOPING CONTAINERS

Think of ADF as the conductor for an orchestra of activities! KNOWS THE BIG PICTURE

2 PIPELINE RUN

AN INSTANCE OF PIPELINE EXECUTION

(arguments) A1 A2 A3 (parameters)

INSTANTIATE PIPELINE BY PASSING ARGUMENTS (values) TO PARAMETERS (placeholders) DEFINED IN IT.

PASS ARGUMENTS MANUALLY OR BY DEFINING TRIGGERS

3 TRIGGERS

TRIGGER EVENT

UNIT OF PROCESSING THAT DETERMINES WHEN TO KICK OFF PIPELINE RUNS.

DIFFERENT TRIGGER TYPES FOR VARIOUS EVENTS!

4 PARAMETERS

"READ-ONLY" KEY/VALUE PAIRS POPULATED FROM RUN CONTEXT AT EXECUTION

BOTH 'DATASET' AND 'LINKED SERVICE' ARE STRONGLY-TYPED, REUSABLE, REFERENCEABLE PARAMETER ENTITIES!!

ACTIVITY REFERENCES 'DATASET' TO CONSUME DATA PROPERTIES, AND 'LINKED SERVICE' FOR THE CONNECTION INFO (to data store for copy, compute resource for flow)

5 VARIABLES

ARE USED INSIDE PIPELINES TO STORE TEMPORARY VALUES, STATE.

USED WITH PARAMETERS TO PASS VALUES BETWEEN ACTIVITIES, DATA FLOWS, PIPELINES.

LET'S TALK ABOUT COMPONENTS OF AZURE DATA FACTORY

HERE ARE THE KEY TERMS AND CONCEPTS YOU NEED TO KNOW TO USE ADF!

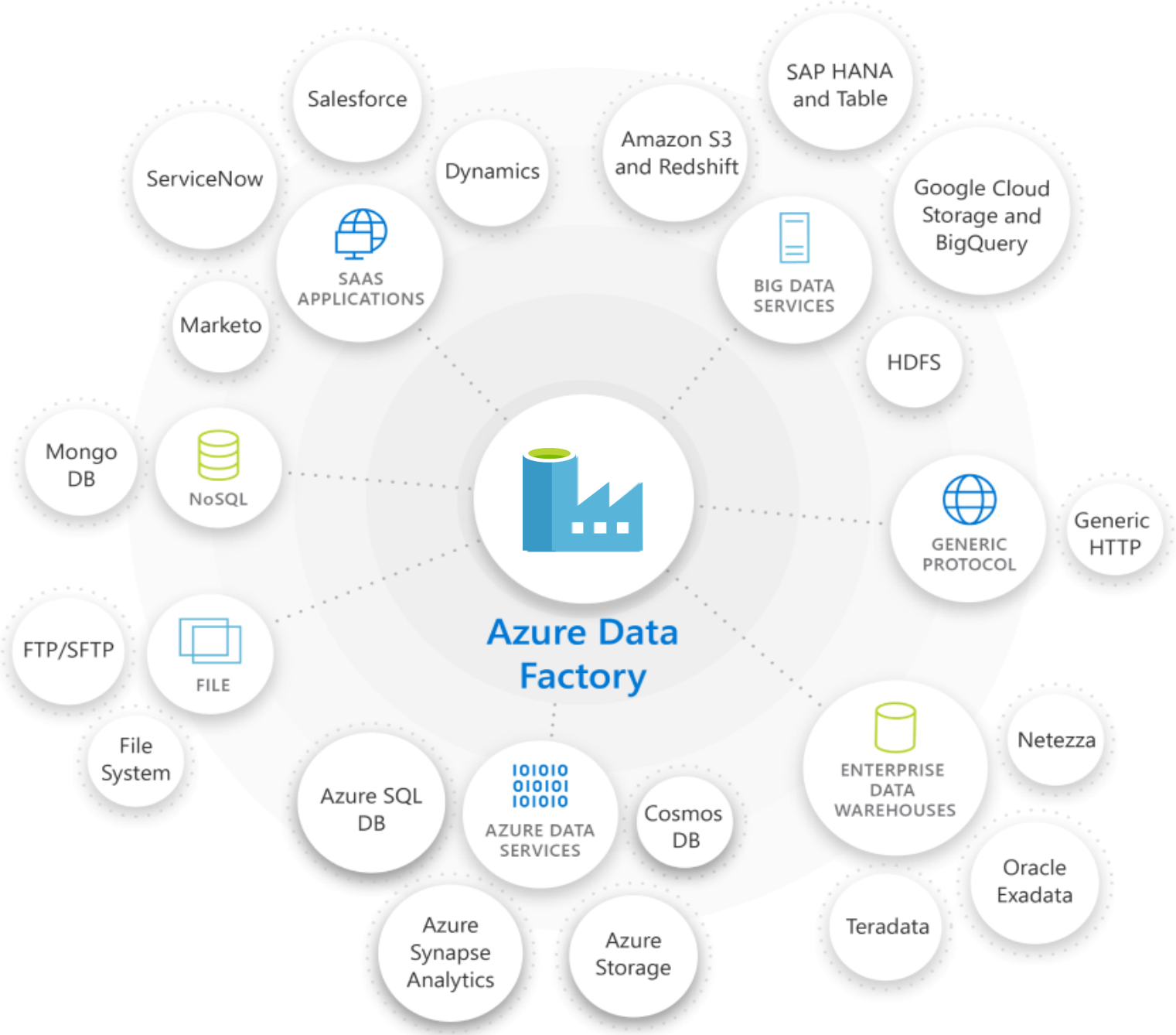
6 MAIN CONCEPTS

- PIPELINES
- ACTIVITIES
- DATASETS
- LINKED SERVICES
- DATA FLOWS
- INTEGRATION RUNTIMES

AND A FEW KEY TERMS

- CONTROL FLOW
- PIPELINE RUN
- TRIGGERS
- PARAMETERS
- VARIABLES

ADF TOOLKIT


Salesforce
ServiceNow
Dynamics
SAAS APPLICATIONS
Marketo
Amazon S3 and Redshift
SAP HANA and Table
Google Cloud Storage and BigQuery
BIG DATA SERVICES
HDFS
Mongo DB
NoSQL
GENERIC PROTOCOL
Generic HTTP
FTP/SFTP
FILE
File System
Azure Data Factory
Azure SQL DB
AZURE DATA SERVICES
Cosmos DB
ENTERPRISE DATA WAREHOUSES
Netezza
Oracle Exadata
Teradata
Azure Synapse Analytics
Azure Storage

CompanyNet

INGEST

- Multi-cloud and on-prem hybrid copy data
- 90+ native connectors
- Serverless and auto-scale
- Use wizard for quick copy jobs

With over 90 native connectors ADF can ingest and transform Data from On Premise File shares, AWS containers, Azure Storage, Oracle databases, SQL and many more.

Data can then be copied and transformed into Azure and on Premise SQL databases , Files shares and even AWS databases for additional reporting or transformation while leaving the source data intact.

Code-Free ETL as a Service

INGEST

- Multi-cloud and on-prem hybrid copy data
- 90+ native connectors
- Serverless and auto-scale
- Use wizard for quick copy jobs

CONTROL FLOW

- Design code-free data pipelines
- Generate pipelines via SDK
- Utilize workflow constructs: loops, branches, conditional execution, variables, parameters, ...

DATA FLOW

- Code-free data transformations that execute in Spark
- Scale-out with Azure Integration Runtimes
- Generate data flows via SDK
- Designers for data engineers and data analysts

SCHEDULE

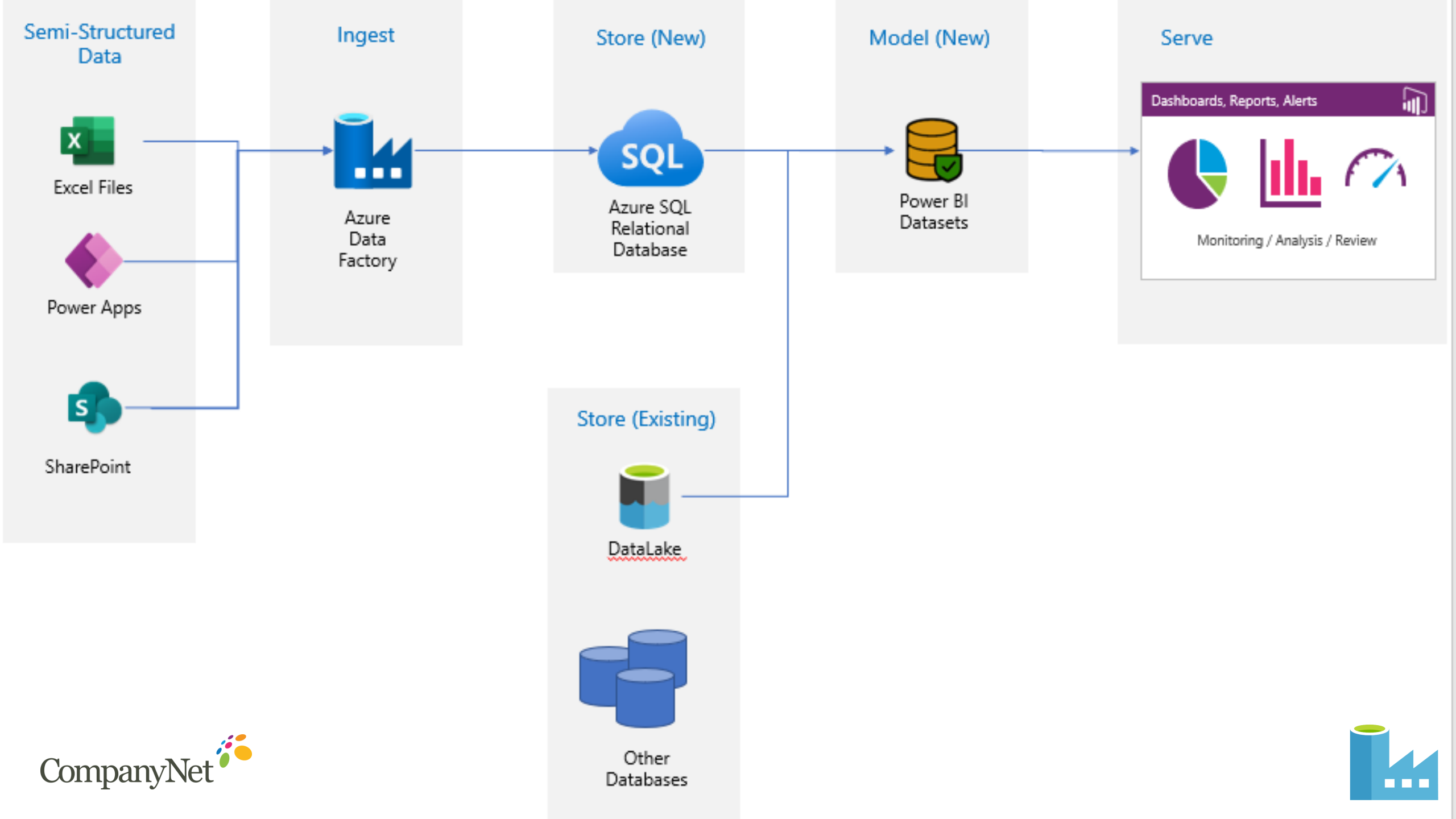
- Build and maintain operational schedules for your data pipelines
- Wall clock, event-based, tumbling windows, chained

MONITOR

- View active executions and pipeline history
- Detail activity and data flow executions
- Establish alerts and notifications

Data Factory allows us to create data-driven workflows in the cloud for orchestrating and automating data movement and data transformation.

CompanyNet

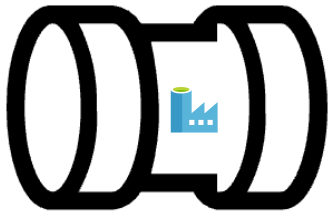

Semi-Structured Data
Excel Files
Power Apps
SharePoint
Ingest
Azure Data Factory
Store (New)
SQL
Azure SQL Relational Database
Model (New)
Power BI Datasets
Serve
Dashboards, Reports, Alerts
Monitoring / Analysis / Review
Store (Existing)
DataLake
Other Databases
CompanyNet

Data Factory **Components**?

Azure Data Factory is composed of these key components

- Pipelines
- Activities
- Datasets
- Linked services
- Data Flows
- Integration Runtimes

These components work together to provide the platform on which you can compose data-driven workflows to move and transform data.



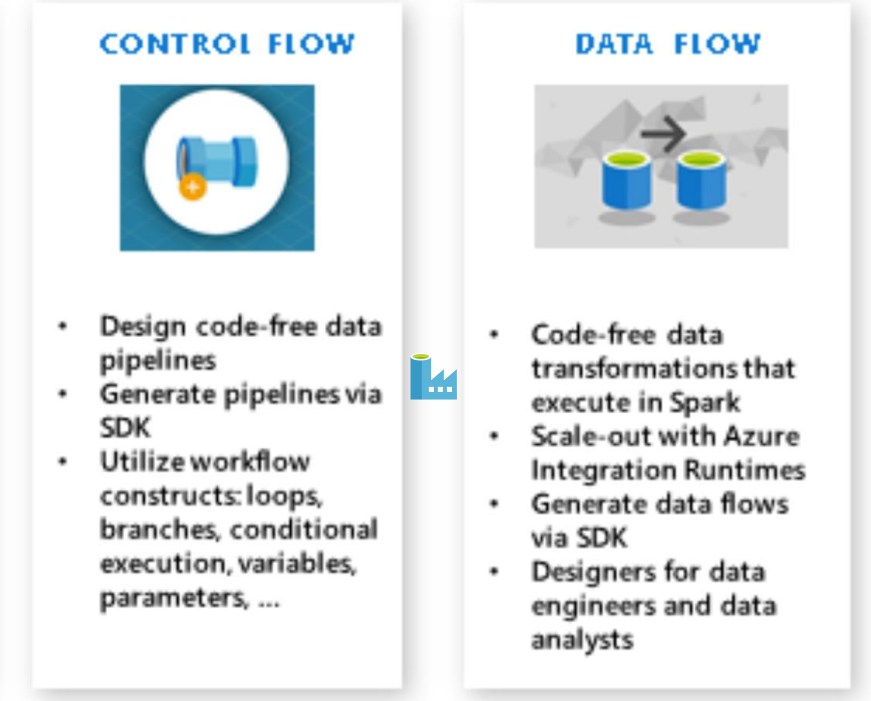
ADF **Pipeline**

A pipeline is a logical grouping of activities that performs a unit of work. Together, the activities in a pipeline perform a task.

The benefit of this is that the pipeline allows you to manage the activities in a set instead of managing each one individually.

The activities in a pipeline can be chained together to operate sequentially, or they can operate independently in parallel all in the same pipeline run.

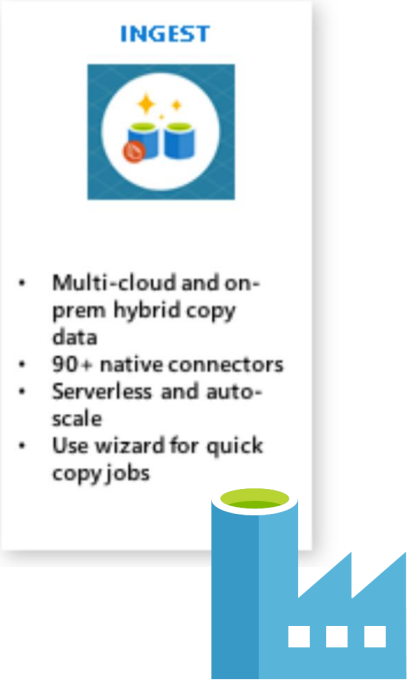
.



The activities in a pipeline define actions to perform on your data.

Pipelines are created code-free using the new V2 Datafactory interface meaning it is easier than ever to create pipelines for Data move and transform activity.

Behind the scenes the managed Apache Spark™ service takes care of the code generation and maintenance.



We will now demo some basic Data Factory functionality in a test environment.

CompanyNet

DEMO

CompanyNet

Thank You!

Q&A