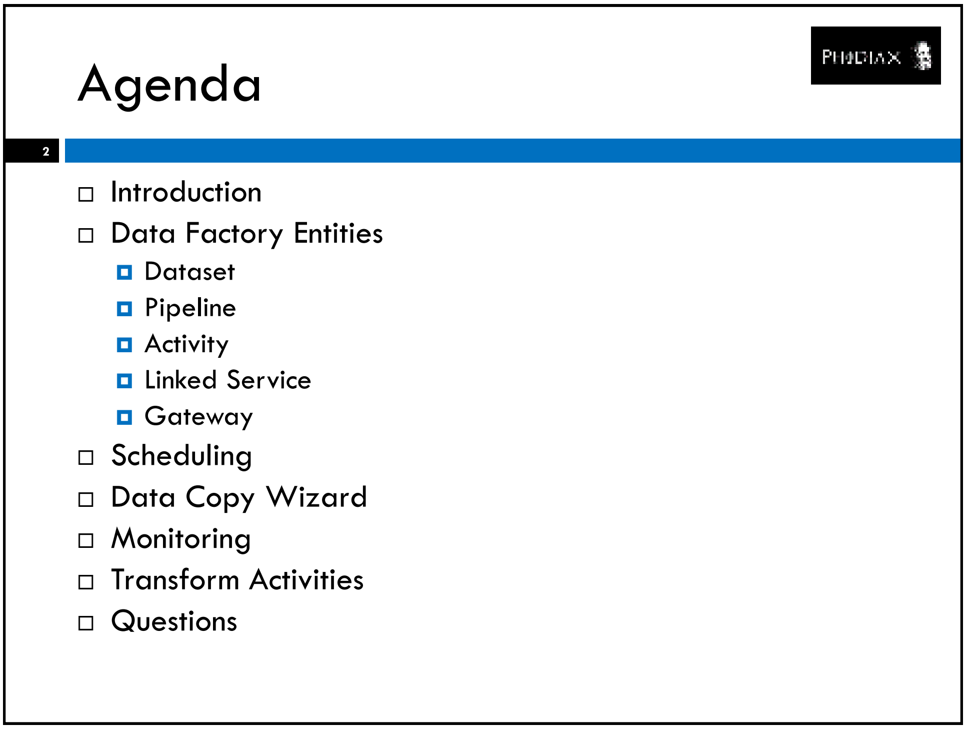
INTRO TO AZURE DATA FACTORY

(FOR SSIS DEVELOPERS)

PHIDIAX, LLC

PHIDIAX

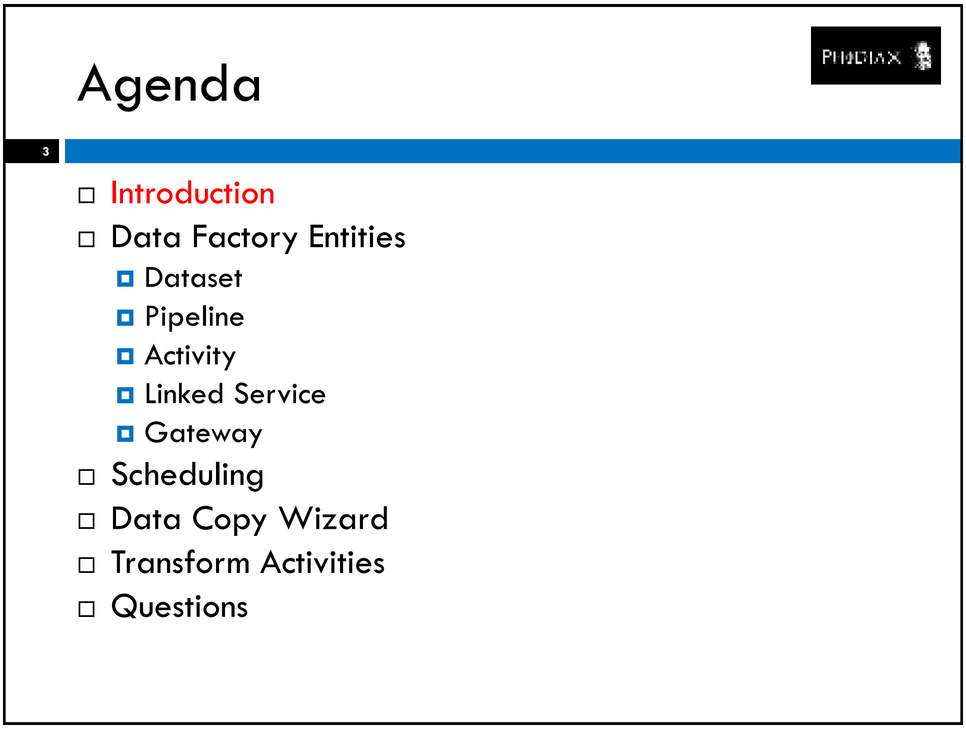
Microsoft Cloud Solution Provider
Managed Gold Partner – System Integrator



Agenda

PHIDIAX

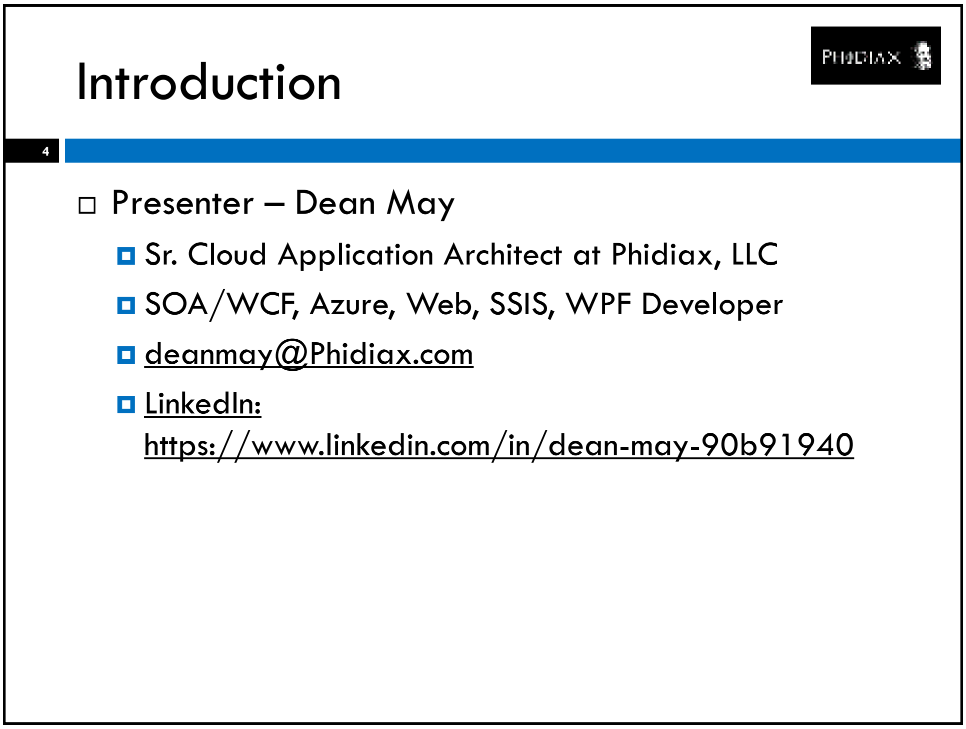
- □ Introduction
- □ Data Factory Entities
  - Dataset
  - Pipeline
  - Activity
  - Linked Service
  - Gateway
- □ Scheduling
- □ Data Copy Wizard
- □ Monitoring
- □ Transform Activities
- □ Questions



Agenda

3

- □ Introduction
- □ Data Factory Entities
  - Dataset
  - Pipeline
  - Activity
  - Linked Service
  - Gateway
- □ Scheduling
- □ Data Copy Wizard
- □ Transform Activities
- □ Questions



Introduction

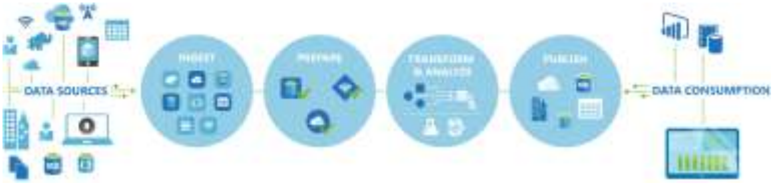
4

- □ Presenter – Dean May
  - Sr. Cloud Application Architect at Phidiax, LLC
  - SOA/WCF, Azure, Web, SSIS, WPF Developer
  - deanmay@Phidiax.com
  - LinkedIn: https://www.linkedin.com/in/dean-may-90b91940

Introduction

5

- □ What is Data Factory?
  - Data Factory is a cloud-based service for moving and transforming data
  - Data Factory provides scheduling and slicing facilities



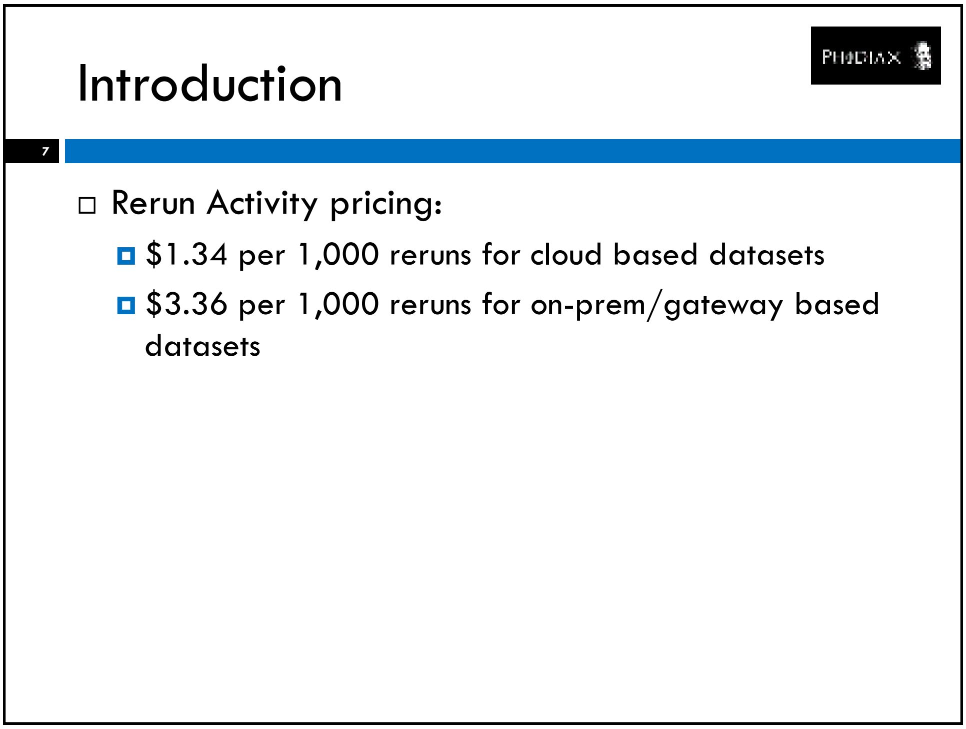
Introduction

6

- □ Data Factory Activity Costing (first five activities are free):

Frequency	Cloud/On-Prem	Activity Cost ($/activity/month)	Data Movement Cost ($/hour)
Daily or less (low)	Cloud	$0.60	$0.25
	On-Prem/Gateway	$1.00	$0.10
More often (high)	Cloud	$1.50	$0.25
	On-Prem/Gateway	$2.50	$0.10
Inactive	Cloud/On-Prem	$0.80 (pipeline/month)	N/A

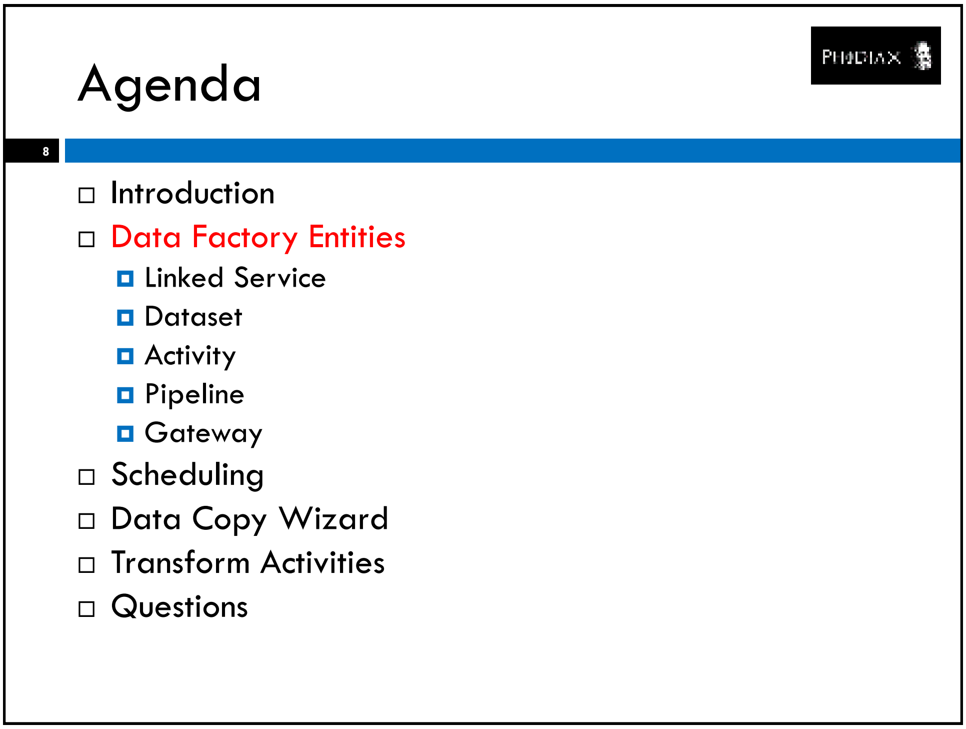
- □ Batch pricing for custom activities based on selected VM sizes to be used



Introduction

7

- □ Rerun Activity pricing:
  - $1.34 per 1,000 reruns for cloud based datasets
  - $3.36 per 1,000 reruns for on-prem/gateway based datasets



Agenda

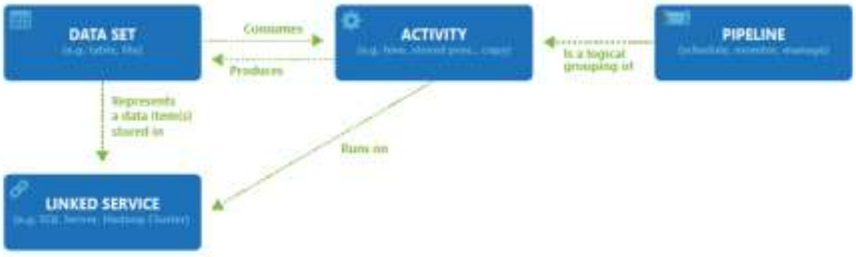
8

- □ Introduction
- □ Data Factory Entities
  - Linked Service
  - Dataset
  - Activity
  - Pipeline
  - Gateway
- □ Scheduling
- □ Data Copy Wizard
- □ Transform Activities
- □ Questions

Data Factory Entities

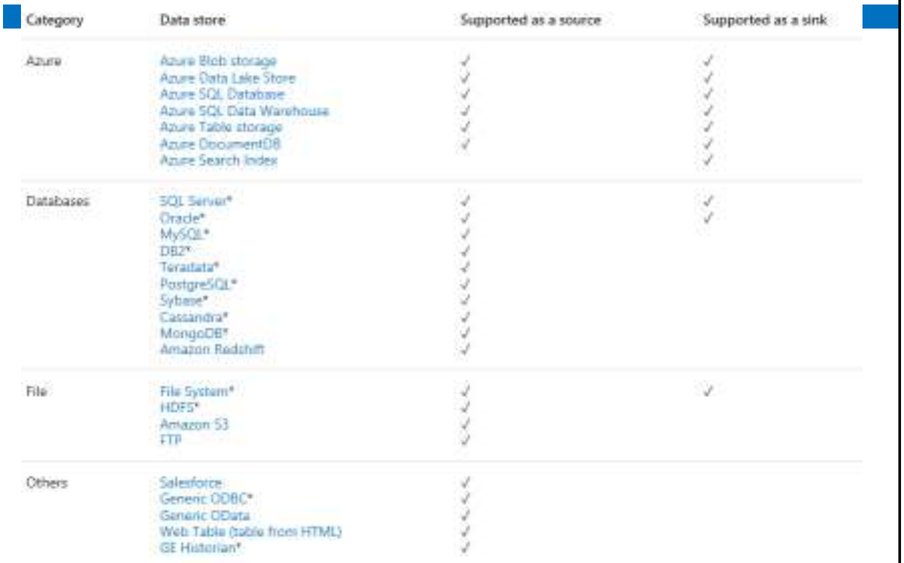
9

- □ Linked Service – a data source (i.e. SQL Server)
- □ Dataset – specific data tables, views, and files
- □ Activity – specific action to take on dataset(s)
- □ Pipeline – logical grouping of activities
- □ Gateway – Software to allow on premise data sources to be consumed and published to by Azure Data Factory

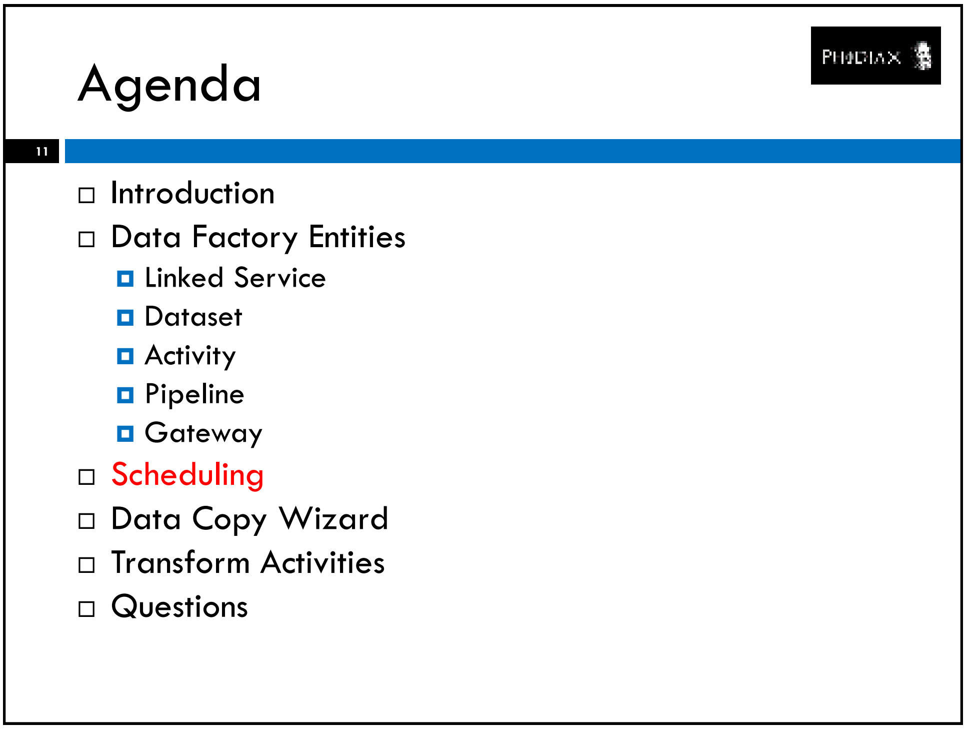

DATA SET
(e.g. table, file)
Consumes
Produces
ACTIVITY
(e.g. hive, stored proc., copy)
Is a logical grouping of
PIPELINE
(schedule, monitor, manage)
Represents a data item(s) stored in
Runs on
LINKED SERVICE
(e.g. SQL Server, Hadoop Cluster)

Data Factory Entities

10



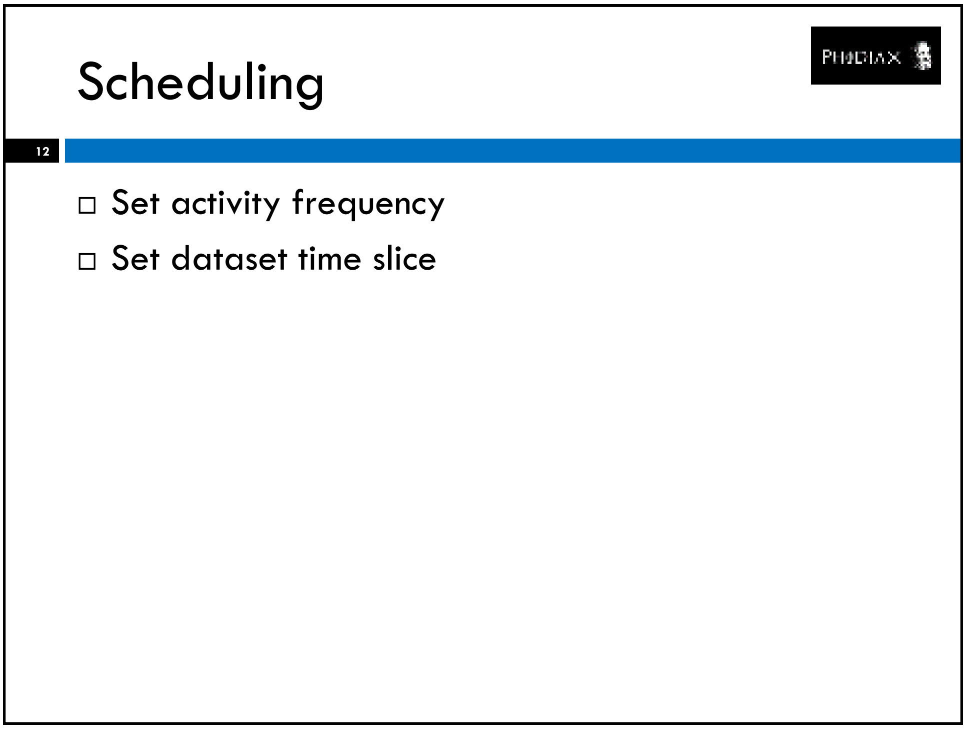
Category	Data store	Supported as a source	Supported as a sink
Azure	Azure Blob storage	✓	✓
	Azure Data Lake Store	✓	✓
	Azure SQL Database	✓	✓
	Azure SQL Data Warehouse	✓	✓
	Azure Table storage	✓	✓
	Azure DocumentDB	✓	✓
	Azure Search Index		✓
Databases	SQL Server*	✓	✓
	Oracle*	✓	✓
	MySQL*	✓	
	DB2*	✓	
	Teradata*	✓	
	PostgreSQL*	✓	
	Sybase*	✓	
	Cassandra*	✓	
	MongoDB*	✓	
	Amazon Redshift	✓	
File	File System*	✓	✓
	HDFS*	✓	
	Amazon S3	✓	
	FTP	✓	
Others	Salesforce	✓	
	Generic ODBC*	✓	
	Generic OData	✓	
	Web Table (table from HTML)	✓	
	GE Historian*	✓	



Agenda

11

- □ Introduction
- □ Data Factory Entities
  - Linked Service
  - Dataset
  - Activity
  - Pipeline
  - Gateway
- □ Scheduling
- □ Data Copy Wizard
- □ Transform Activities
- □ Questions



Scheduling

12

- □ Set activity frequency
- □ Set dataset time slice

Agenda

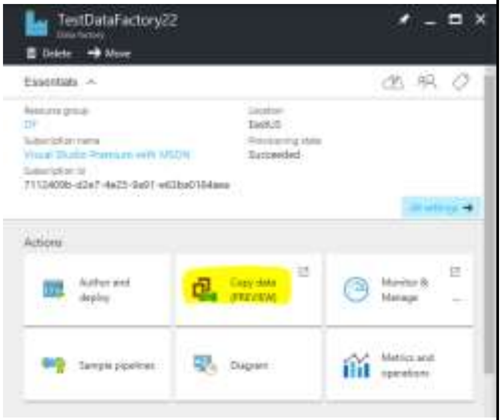
13

- □ Introduction
- □ Data Factory Entities
  - Linked Service
  - Dataset
  - Activity
  - Pipeline
  - Gateway
- □ Scheduling
- □ Data Copy Wizard
- □ Transform Activities
- □ Questions

Data Copy Wizard

14

- □ The data copy wizard provides a web interface to facilitate creating input data source, an activity/pipeline, and an output data source.
  - Sample Azure SQL to Azure Blob File



TestDataFactory22

Resource group
DF

Location
EastUS

Subscription name
Visual Studio Premium with MSDN

Provisioning state
Succeeded

Subscription id
71124000b-d2e7-4e25-9a91-e63ba0184aea

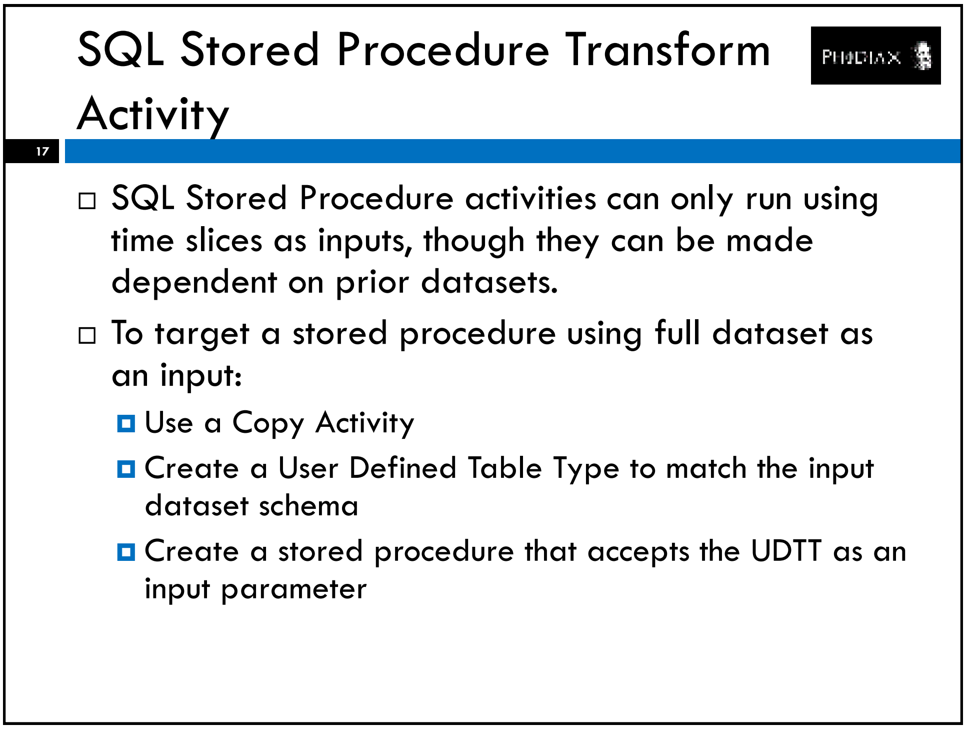
15

Demo Data Copy Wizard / Monitoring

Agenda

16

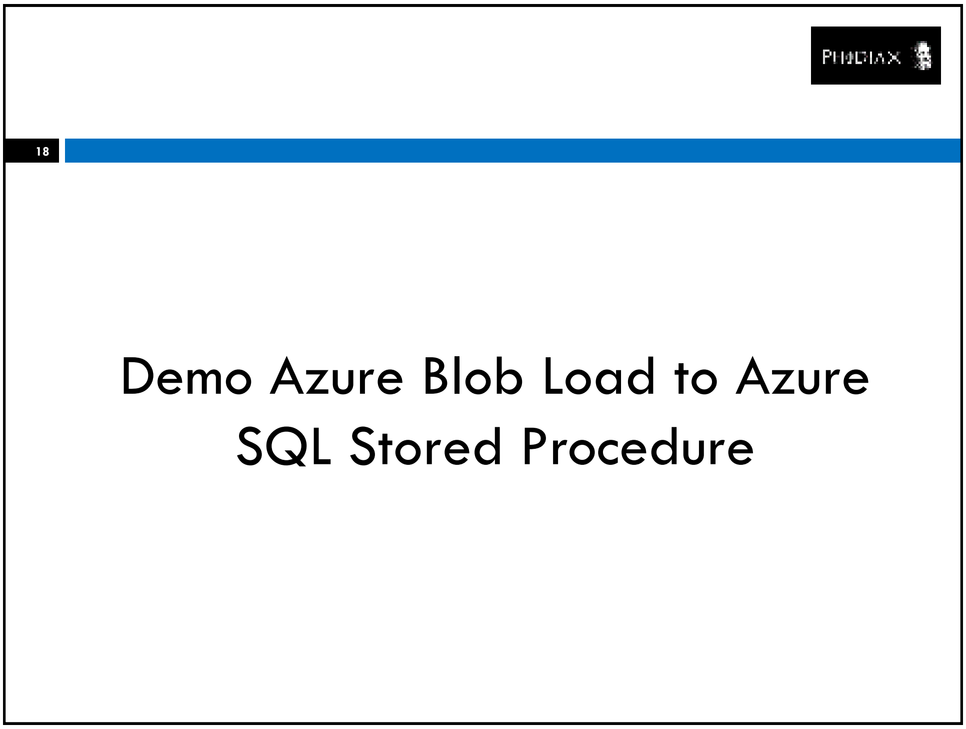
- □ Introduction
- □ Data Factory Entities
  - Linked Service
  - Dataset
  - Activity
  - Pipeline
  - Gateway
- □ Scheduling
- □ Data Copy Wizard
- □ Transform Activities
- □ Questions



SQL Stored Procedure Transform Activity

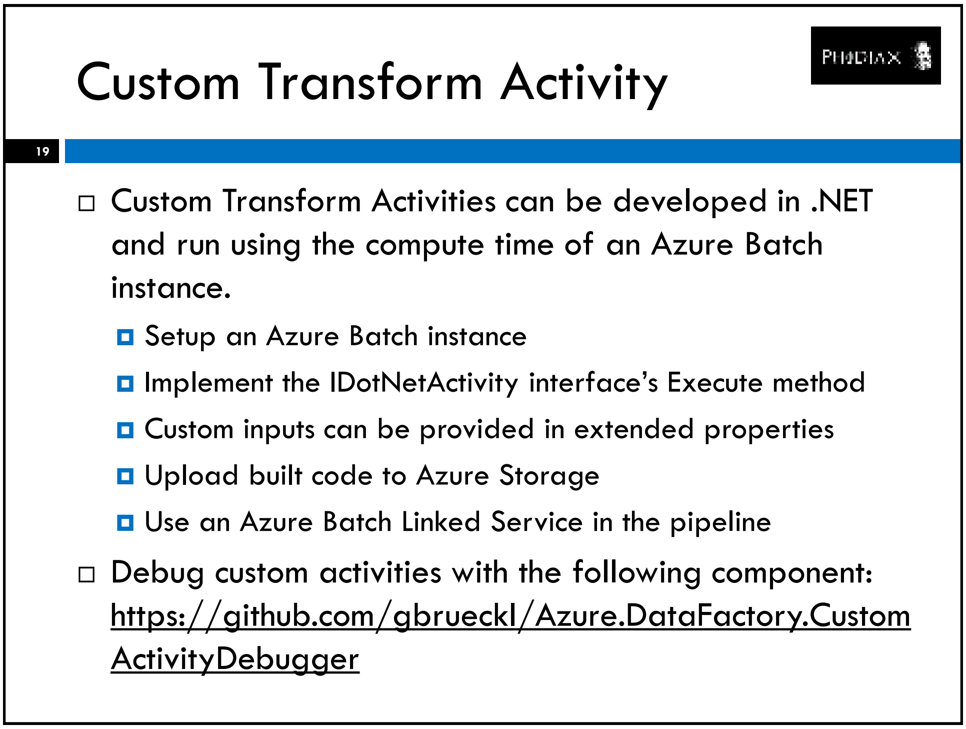
17

- SQL Stored Procedure activities can only run using time slices as inputs, though they can be made dependent on prior datasets.
- To target a stored procedure using full dataset as an input:
  - Use a Copy Activity
  - Create a User Defined Table Type to match the input dataset schema
  - Create a stored procedure that accepts the UDTT as an input parameter



18

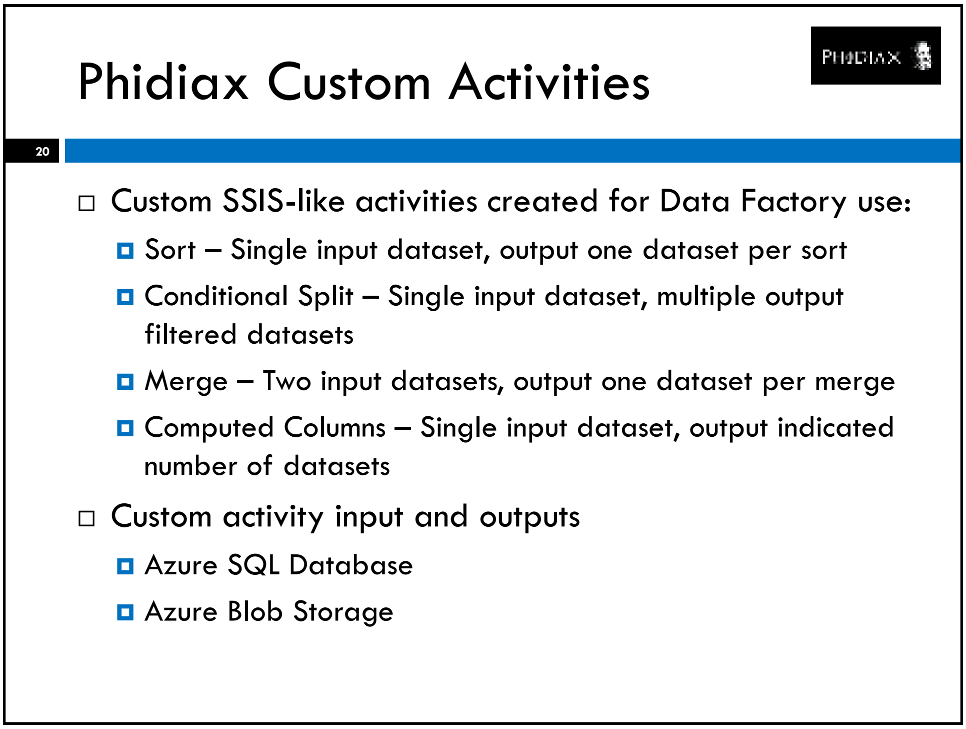
Demo Azure Blob Load to Azure SQL Stored Procedure



Custom Transform Activity

19

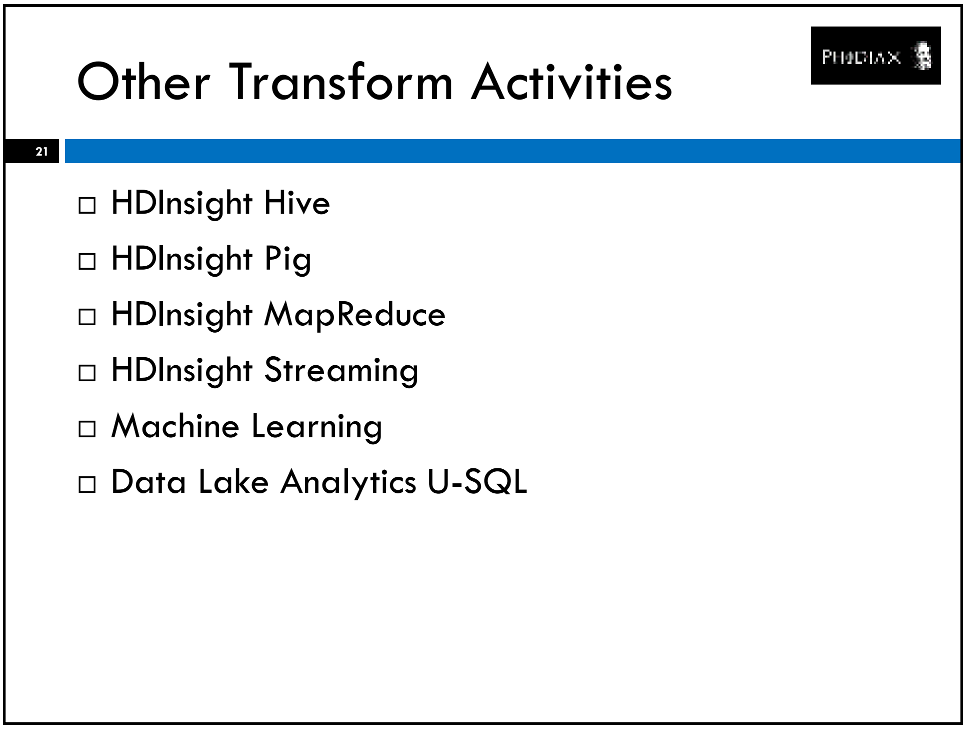
- Custom Transform Activities can be developed in .NET and run using the compute time of an Azure Batch instance.
  - Setup an Azure Batch instance
  - Implement the IDotNetActivity interface's Execute method
  - Custom inputs can be provided in extended properties
  - Upload built code to Azure Storage
  - Use an Azure Batch Linked Service in the pipeline
- Debug custom activities with the following component: https://github.com/gbrueckl/Azure.DataFactory.CustomActivityDebugger



Phidiax Custom Activities

20

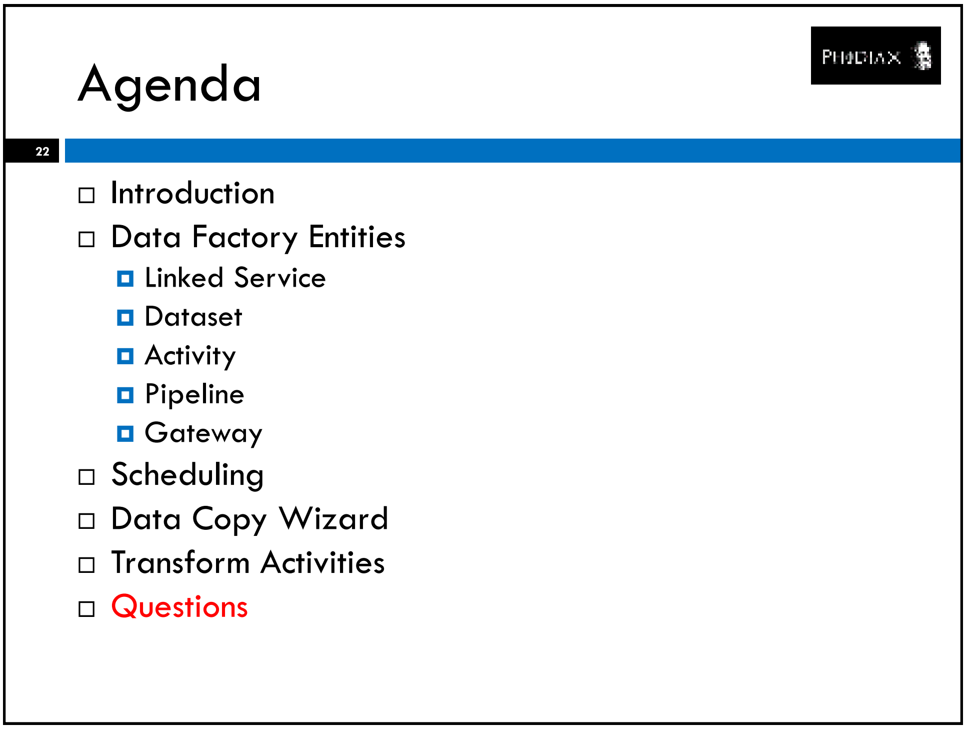
- Custom SSIS-like activities created for Data Factory use:
  - Sort – Single input dataset, output one dataset per sort
  - Conditional Split – Single input dataset, multiple output filtered datasets
  - Merge – Two input datasets, output one dataset per merge
  - Computed Columns – Single input dataset, output indicated number of datasets
- Custom activity input and outputs
  - Azure SQL Database
  - Azure Blob Storage



Other Transform Activities

21

- □ HDInsight Hive
- □ HDInsight Pig
- □ HDInsight MapReduce
- □ HDInsight Streaming
- □ Machine Learning
- □ Data Lake Analytics U-SQL



Agenda

22

- □ Introduction
- □ Data Factory Entities
  - □ Linked Service
  - □ Dataset
  - □ Activity
  - □ Pipeline
  - □ Gateway
- □ Scheduling
- □ Data Copy Wizard
- □ Transform Activities
- □ Questions

23

Demo & Questions?