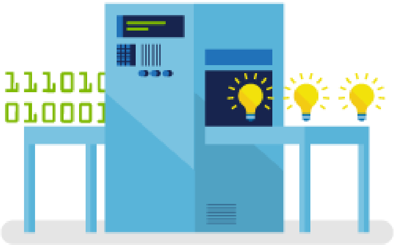
Microsoft

ELT using Azure Databricks and Data Factory

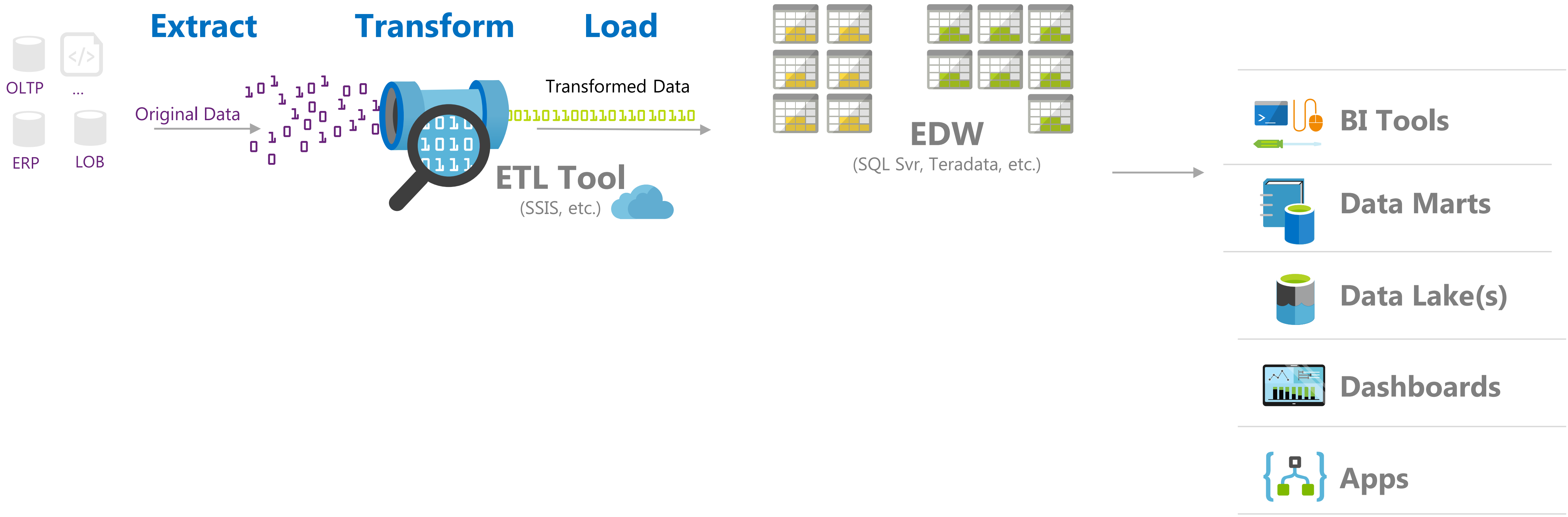
Gaurav Malhotra
Senior Program Manager-Microsoft

Agenda



- Modern Data Engineering
- Azure Data Factory Overview
- Azure Databricks Overview
- Demos
- Q & A

Modern Data Engineering



Extract

Transform

Load

OLTP

...

ERP

LOB

Original Data

Transformed Data

ETL Tool

(SSIS, etc.)

EDW

(SQL Svr, Teradata, etc.)

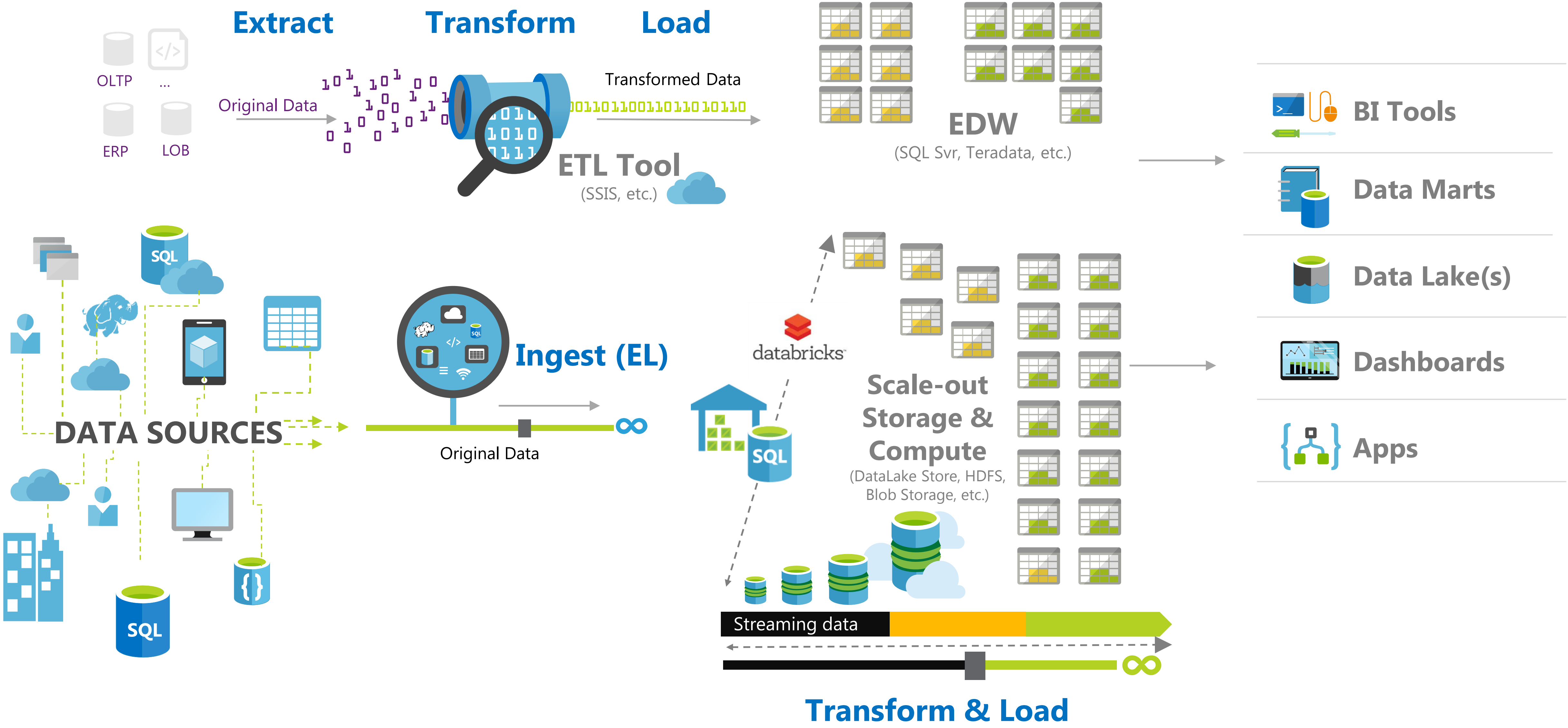
BI Tools

Data Marts

Data Lake(s)

Dashboards

Apps



Extract Transform Load

OLTP ... ERP LOB

Original Data

Transformed Data

ETL Tool

(SSIS, etc.)

EDW

(SQL Svr, Teradata, etc.)

DATA SOURCES

Ingest (EL)

Original Data

databricks

Scale-out Storage & Compute

(DataLake Store, HDFS, Blob Storage, etc.)

Streaming data

Transform & Load

BI Tools

Data Marts

Data Lake(s)

Dashboards

Apps

Azure Data Factory

Managed Data Integration Service

Azure Data Factory

Managed Data Integration Service

Flexible Pipeline Model

Rich pipeline orchestration

Triggers: on-demand, schedule, event

Data Movement as a Service

Cloud, Hybrid

70+ connectors provided

SSIS Package Execution

In a managed cloud environment

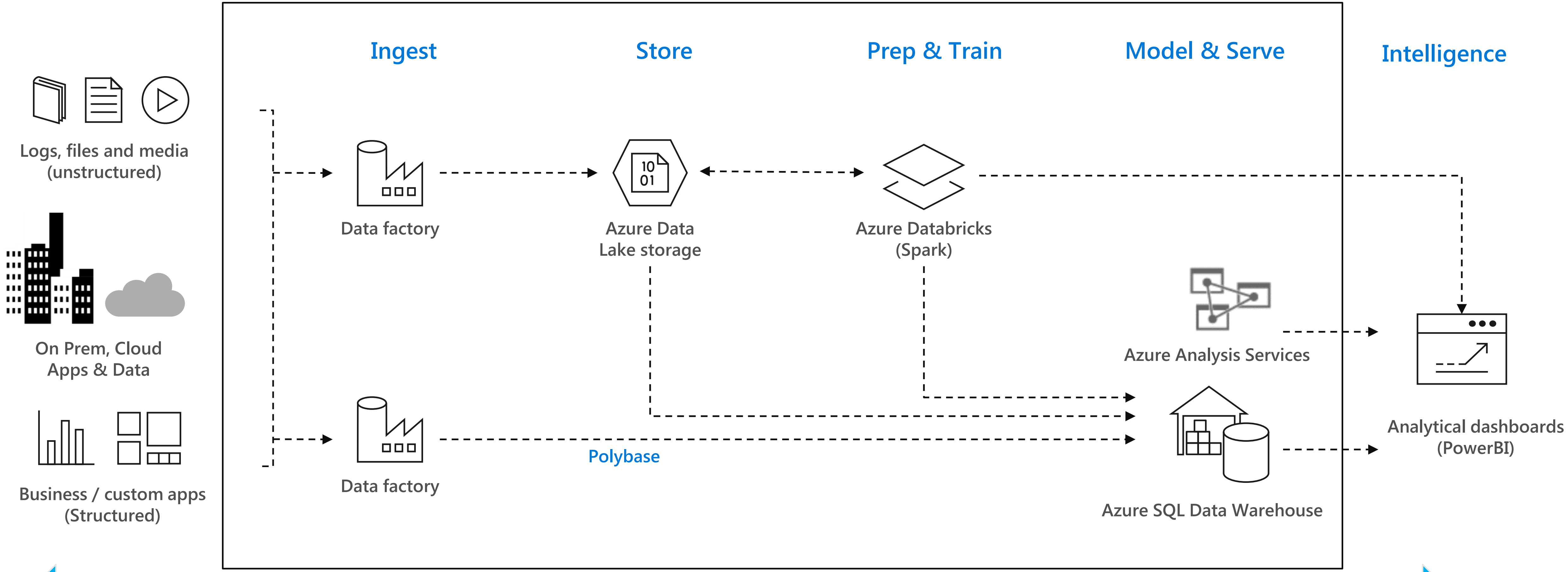
Use familiar tools, SSMS & SSDT

Author & Monitor

Programmability (Python, .NET, Powershell, etc.)

Visual Tools

Modern Data Engineering for BI



Ingest

Store

Prep & Train

Model & Serve

Intelligence

Logs, files and media
(unstructured)

On Prem, Cloud
Apps & Data

Business / custom apps
(Structured)

Data factory

Azure Data
Lake storage

Azure Databricks
(Spark)

Azure Analysis Services

Data factory

Polybase

Azure SQL Data Warehouse

Analytical dashboards
(PowerBI)

AZURE DATA FACTORY ORCHESTRATES DATA PIPELINE ACTIVITY WORKFLOW & SCHEDULING

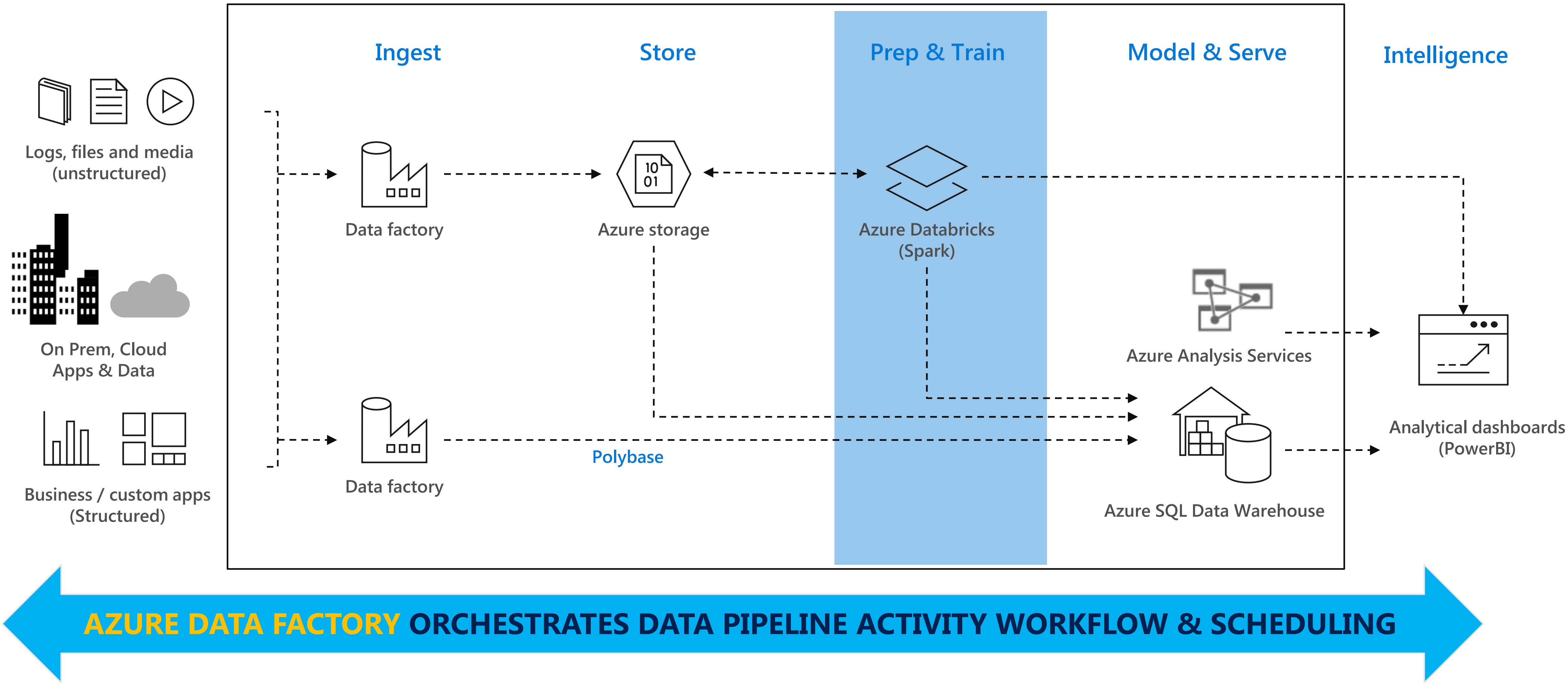
Azure Databricks

Trusted & Reliable Platform for Data Engineering

AZURE DATABRICKS

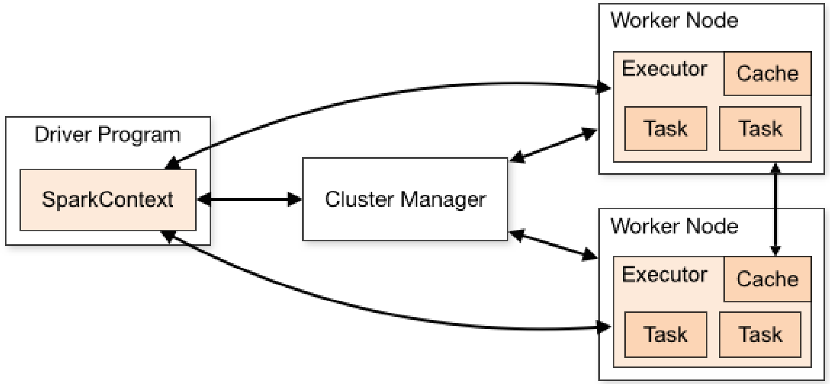
- Azure Databricks is a **first party** service on Azure.
  - Unlike with other clouds, it is not an Azure Marketplace or a 3rd party hosted service.
- Azure Databricks is integrated seamlessly with Azure services:
  - Azure Portal: Service an be launched directly from Azure Portal
  - Azure Storage Services: Directly access data in Azure Blob Storage and Azure Data Lake Store
  - Azure Active Directory: For user authentication, eliminating the need to maintain two separate sets of uses in Databricks and Azure.
  - Azure SQL DW and Azure Cosmos DB: Enables you to combine structured and unstructured data for analytics
  - Apache Kafka for HDInsight: Enables you to use Kafka as a streaming data source or sink
  - Azure Event Hub & Azure IOT Hubs: Enables you to use Event Hub and IOT Hub as a streaming data source
  - Azure Billing: You get a single bill from Azure
  - Azure Power BI: For rich data visualization
  - Azure Data Factory: ETL/ELT - See here

Azure Databricks Role in Modern Data Warehouse

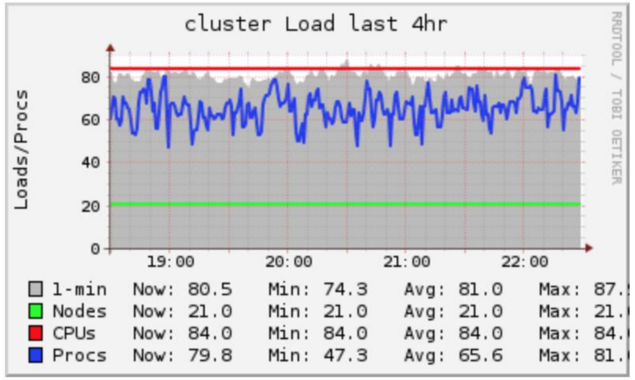


AZURE DATA FACTORY ORCHESTRATES DATA PIPELINE ACTIVITY WORKFLOW & SCHEDULING

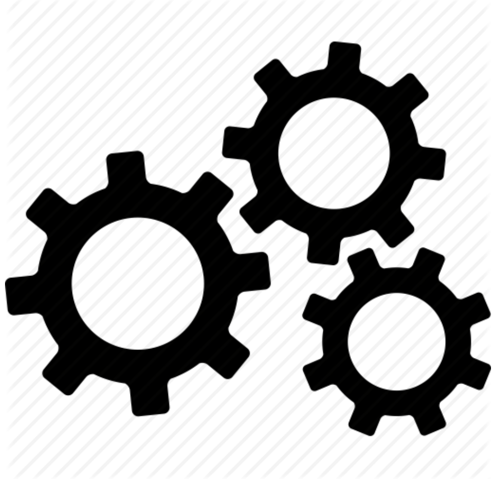
Infinite Scale, Lower Cost, Zero Management



1 to 1000s of Worker Nodes



Auto-scale
Compute & Storage



Auto-Recovery & Upgrade

Your Language, Your Data (Anywhere), Your Format

- SQL, Python, Scala & R Support
  - Code in your favorite language

- Source data from File System, Object stores, HDFS, Database, Pub-Sub systems & Others
  - Read and write data from/to multiple sources
  - Optimized for Azure Blob Store, ADLS, SQLDW, Event Hubs & Cosmos DB

- File Formats
  - CSV, JSON, Parquet, Text, ORC, XML & More

Batch & Streaming Using Unified API

- Structured Streaming
  - Built on Spark SQL Engine
  - Express Streaming computation like batch computation on static data
  - Micro-batch & continuous processing support
  - Fault Tolerant, Only once computation
  - Supports
    - Late Data / Out of Order Data
    - Data de-duplication
    - Stream to Static Join
    - Stream to Stream Join

Azure Databricks Unified Computing simplifies and accelerates Data Engineering

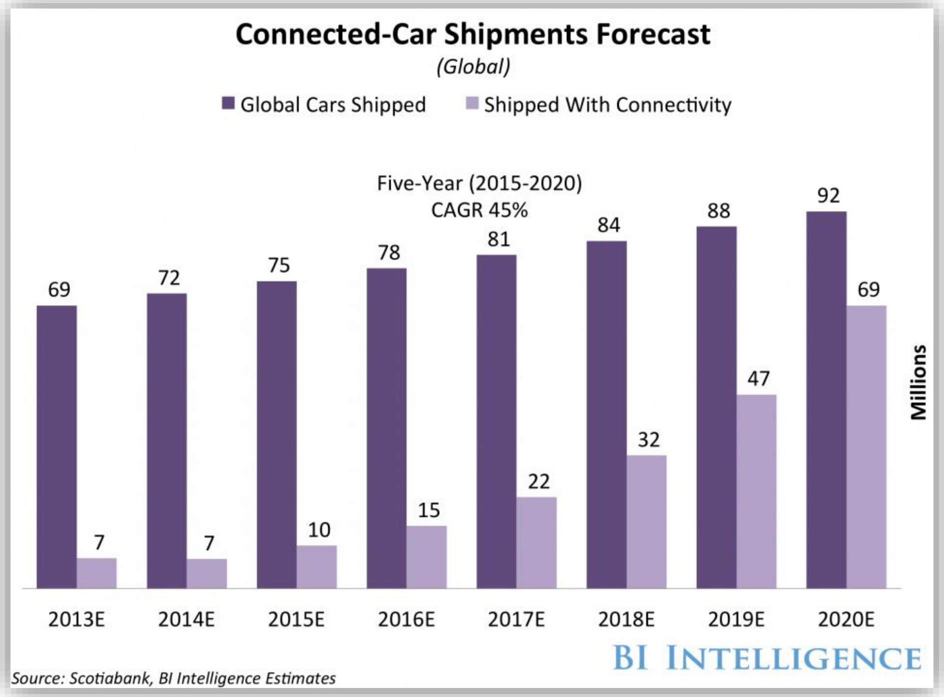
Demo-Connected Cars

Connected car market

- The connected car market is growing
  - **45%** compound annual growth rate over 5 years
  - **10x faster** than overall car market
  - **75%** of cars shipped globally by **2020** will have necessary hardware to connect to the **Internet**

- Connected car technology is split between two approaches
  - Put the Internet connection in the car (**embedded connections**)
    - Does not require a phone data plan to operate
    - Provides access to more features and data
  - Rely on a **secondary device**

- **Embedded connections win**, because auto companies will be able to
  - Collect data on the performance of cars
  - Send updates and patches to cars remotely
  - Avoid recalls related to the car's software



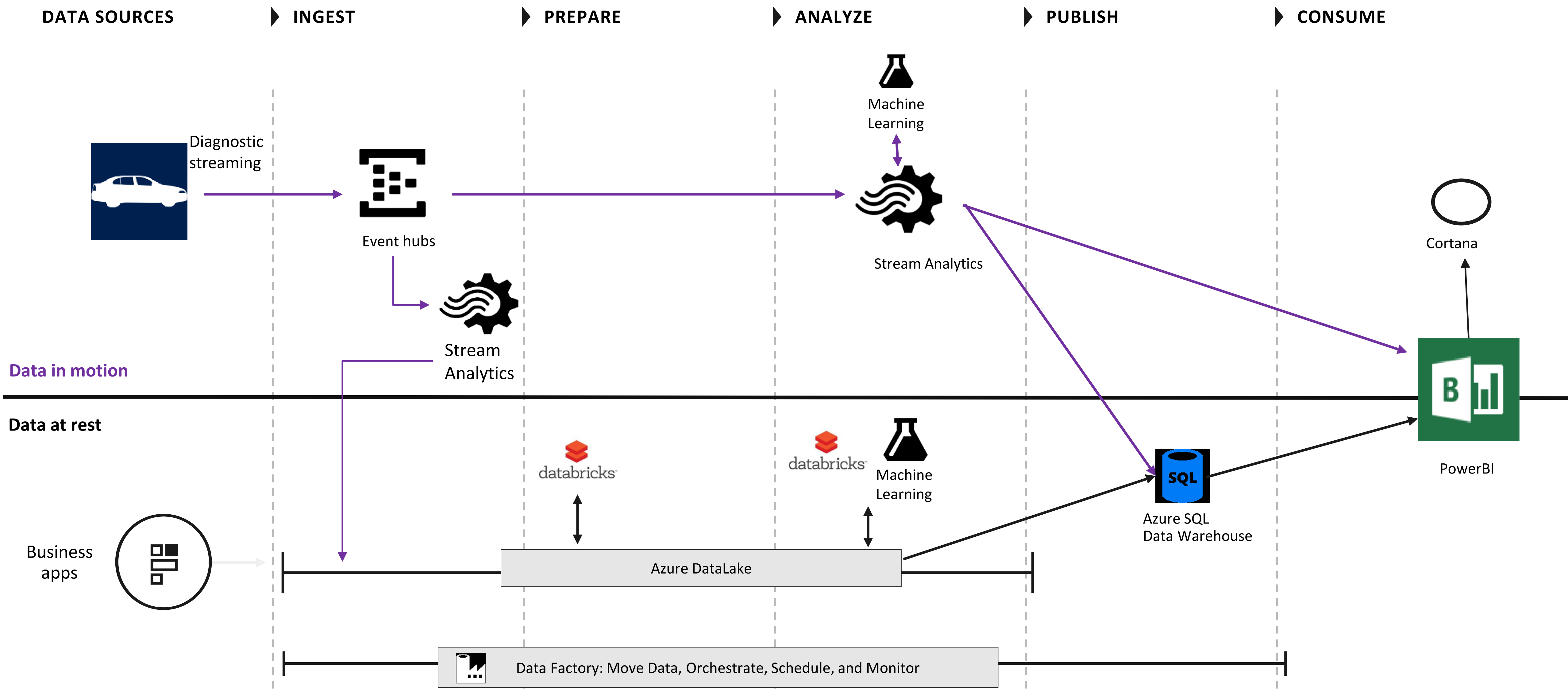
Source: Scotiabank, BI Intelligence Estimates

Connected car market

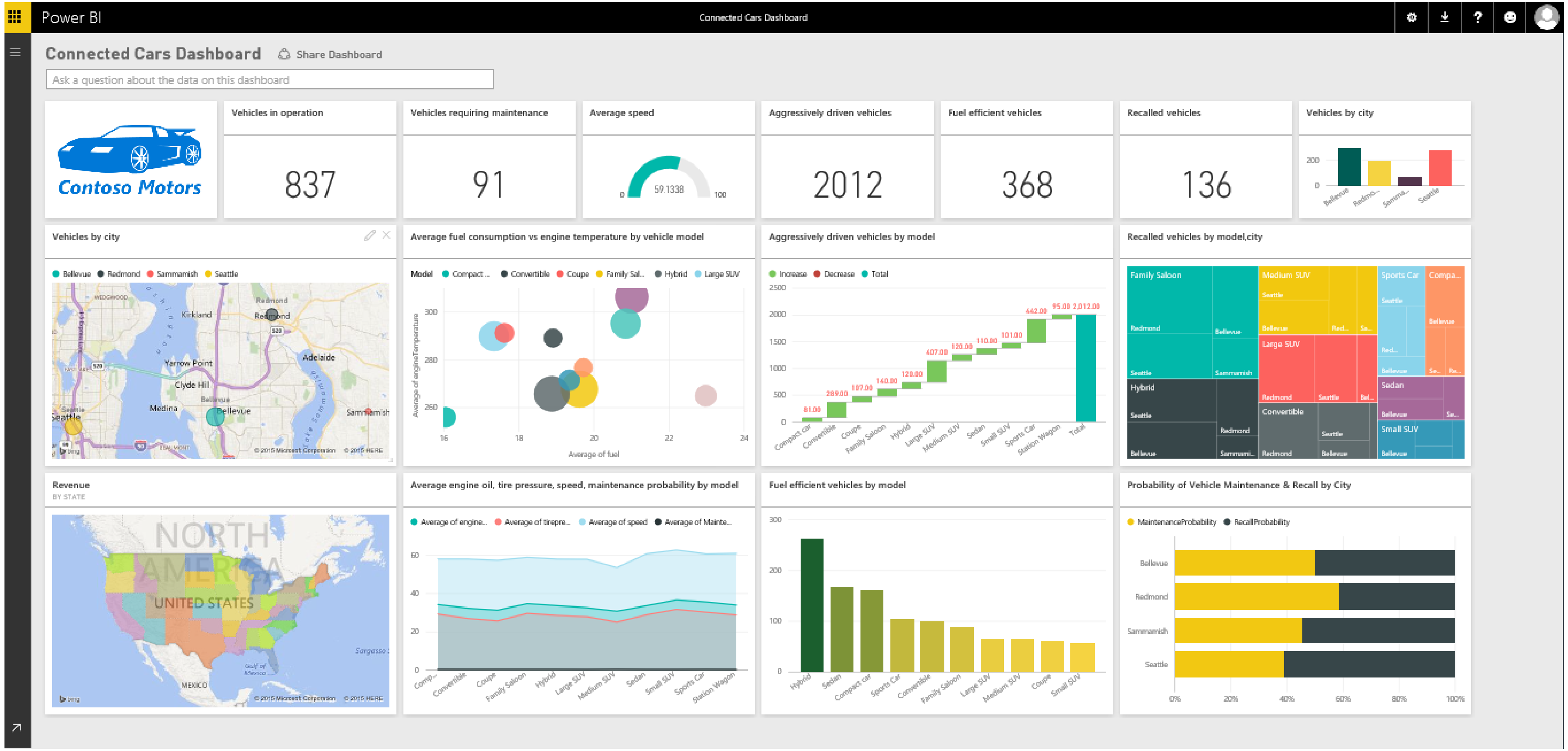
75% of the cars shipped globally by **2020** will be built with the necessary hardware to connect to the Internet

- Vehicle diagnostic
- Usage-based insurance
- Fleet management
- Roadside assistance
- Eco-driving
- Engine performance remapping
- Engine emission control

Demo Architecture Diagram-Connected Cars



Power BI dashboard



Power BI

Connected Cars Dashboard

Connected Cars Dashboard

Share Dashboard

Ask a question about the data on this dashboard

Contoso Motors

Vehicles in operation

837

Vehicles requiring maintenance

91

Average speed

59.1338

Aggressively driven vehicles

2012

Fuel efficient vehicles

368

Recalled vehicles

136

Vehicles by city

Vehicles by city

Average fuel consumption vs engine temperature by vehicle model

Aggressively driven vehicles by model

Recalled vehicles by model,city

Revenue

BY STATE

Average engine oil, tire pressure, speed, maintenance probability by model

Fuel efficient vehicles by model

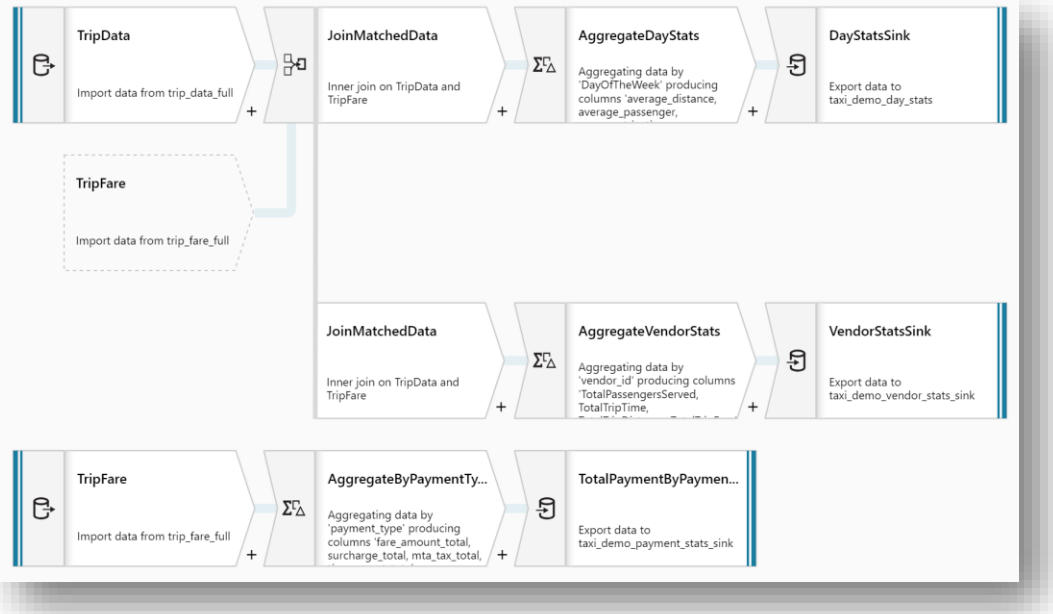
Probability of Vehicle Maintenance & Recall by City

Dataflow

Visual Data Transformation in ADF (Private Preview)

Code-free Data Transformation At Scale

- Does not require understanding of Spark, Big Data Execution Engines, Clusters, Scala …
- Focus on building business logic and data transformation
  - Data cleansing
  - Aggregation
  - Data conversions
  - Data prep
  - Data exploration
  - ETL Data Loading into DW



TripData
Import data from trip_data_full

TripFare
Import data from trip_fare_full

JoinMatchedData
Inner join on TripData and TripFare

AggregateDayStats
Aggregating data by 'DayOfTheWeek' producing columns 'average_distance, average_passenger,

DayStatsSink
Export data to taxi_demo_day_stats

JoinMatchedData
Inner join on TripData and TripFare

AggregateVendorStats
Aggregating data by 'vendor_id' producing columns 'TotalPassengersServed, TotalTripTime,

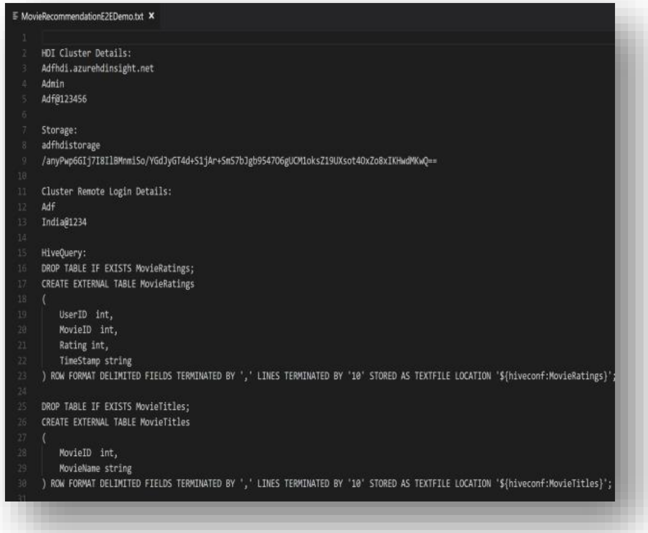
VendorStatsSink
Export data to taxi_demo_vendor_stats_sink

TripFare
Import data from trip_fare_full

AggregateByPaymentTy...
Aggregating data by 'payment_type' producing columns 'fare_amount_total, surcharge_total, mta_tax_total,

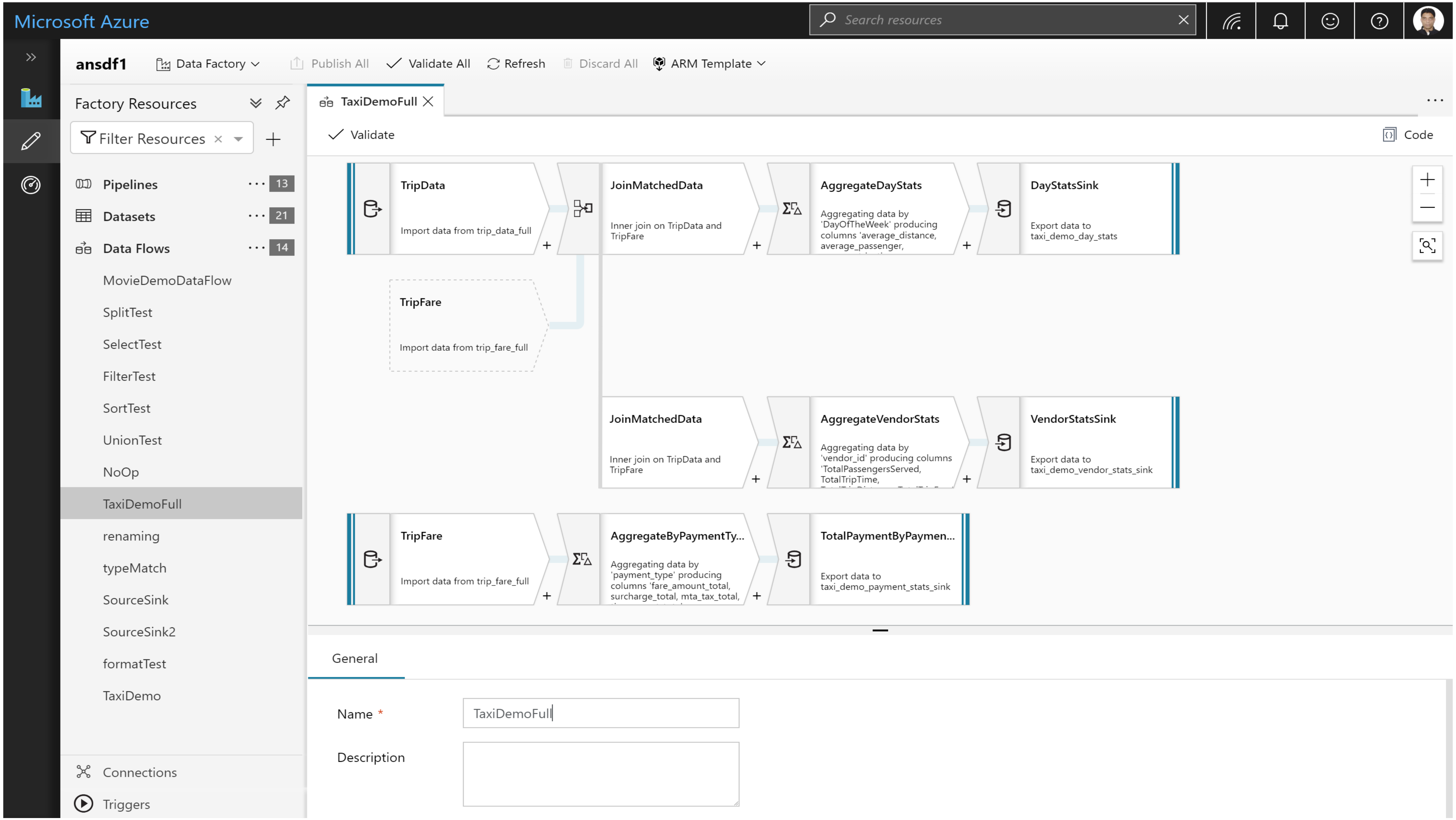
TotalPaymentByPaymen...
Export data to taxi_demo_payment_stats_sink

… not



```

HDI Cluster Details:
Adfhdi.azurehdinsight.net
Admin
Adf@123456

Storage:
adfhdistorage
/anyPwp6GIj7I8IlBMnmiSo/YGdJyGT4d+S1jAr+SmS7bJgb954706gUCMloksZ19UXsot40xZo8xIKHwdMKwQ==

Cluster Remote Login Details:
Adf
India@1234

HiveQuery:
DROP TABLE IF EXISTS MovieRatings;
CREATE EXTERNAL TABLE MovieRatings
(
    UserID  int,
    MovieID  int,
    Rating int,
    TimeStamp string
) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '10' STORED AS TEXTFILE LOCATION '${hiveconf:MovieRatings}';

DROP TABLE IF EXISTS MovieTitles;
CREATE EXTERNAL TABLE MovieTitles
(
    MovieID  int,
    MovieName string
) ROW FORMAT DELIMITED FIELDS TERMINATED BY ',' LINES TERMINATED BY '10' STORED AS TEXTFILE LOCATION '${hiveconf:MovieTitles}';
```

Azure Data Factory Visual Data Flow



TripData
Import data from trip_data_full

JoinMatchedData
Inner join on TripData and TripFare

AggregateDayStats
Aggregating data by 'DayOfTheWeek' producing columns 'average_distance, average_passenger,

DayStatsSink
Export data to taxi_demo_day_stats

TripFare
Import data from trip_fare_full

JoinMatchedData
Inner join on TripData and TripFare

AggregateVendorStats
Aggregating data by 'vendor_id' producing columns 'TotalPassengersServed, TotalTripTime,

VendorStatsSink
Export data to taxi_demo_vendor_stats_sink

TripFare
Import data from trip_fare_full

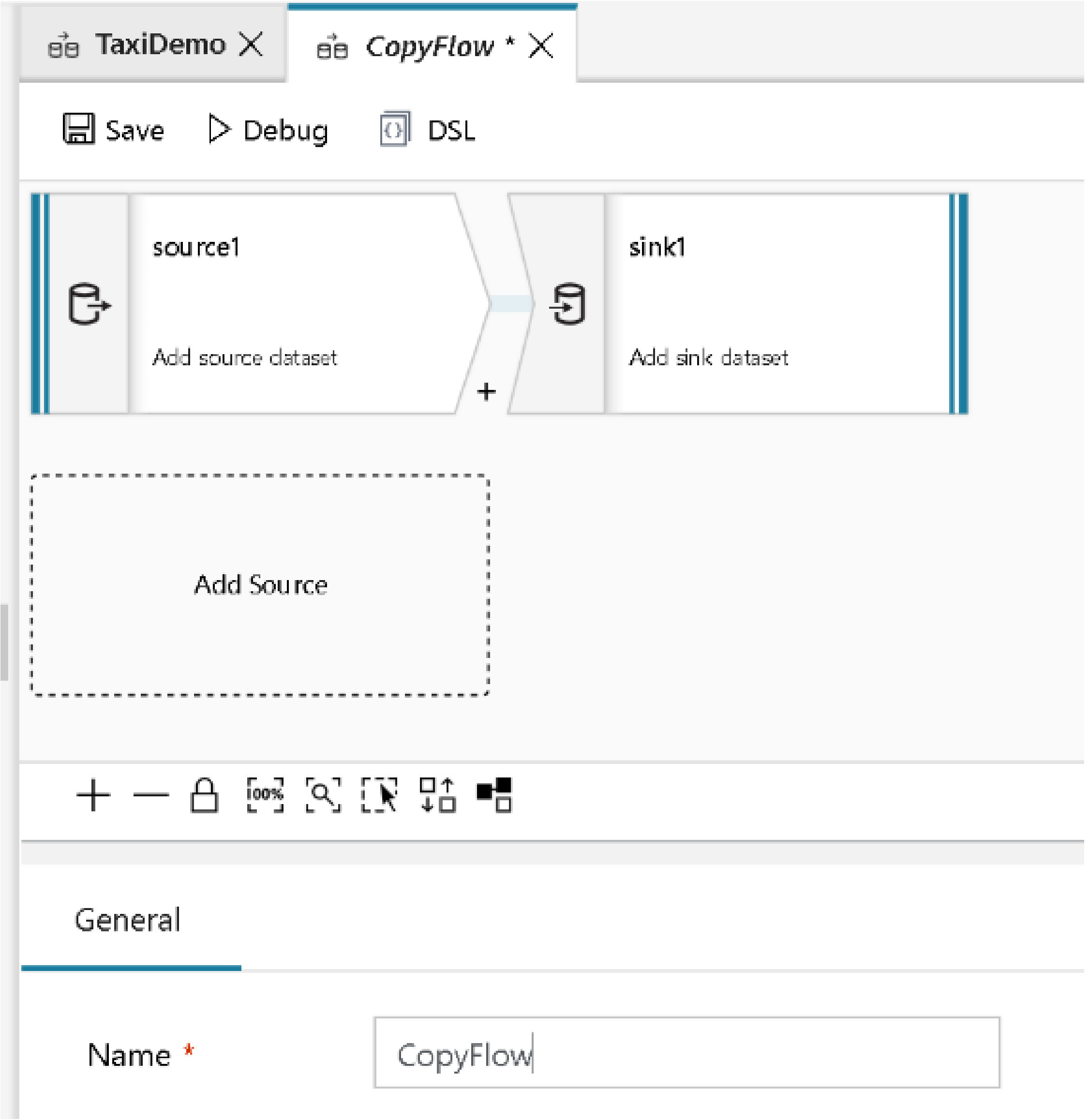
AggregateByPaymentTy...
Aggregating data by 'payment_type' producing columns 'fare_amount_total, surcharge_total, mta_tax_total,

TotalPaymentByPaymen...
Export data to taxi_demo_payment_stats_sink

Name * TaxiDemoFull

Description

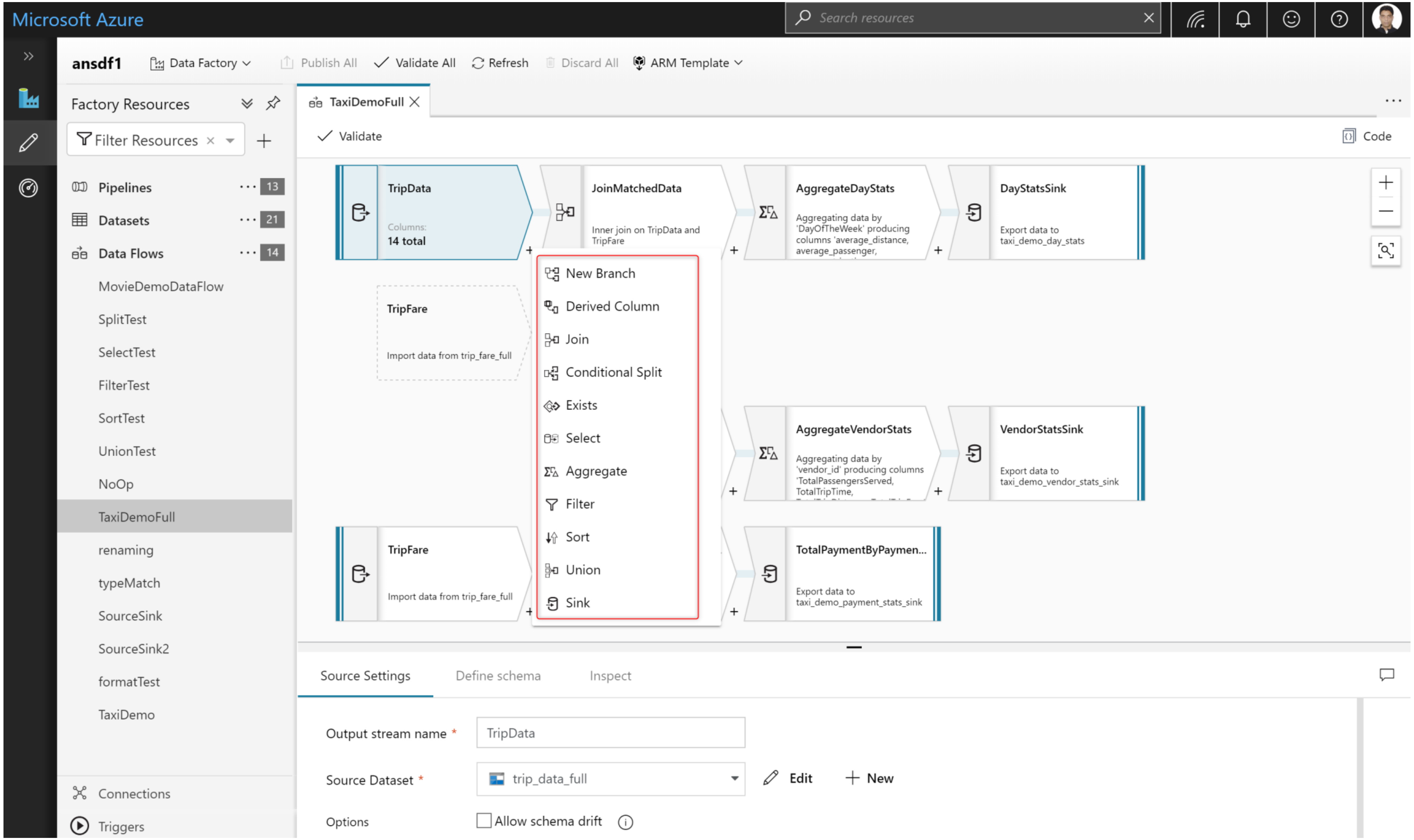
Simple Copy Flow



Name *

CopyFlow

Guided experience to build data flows

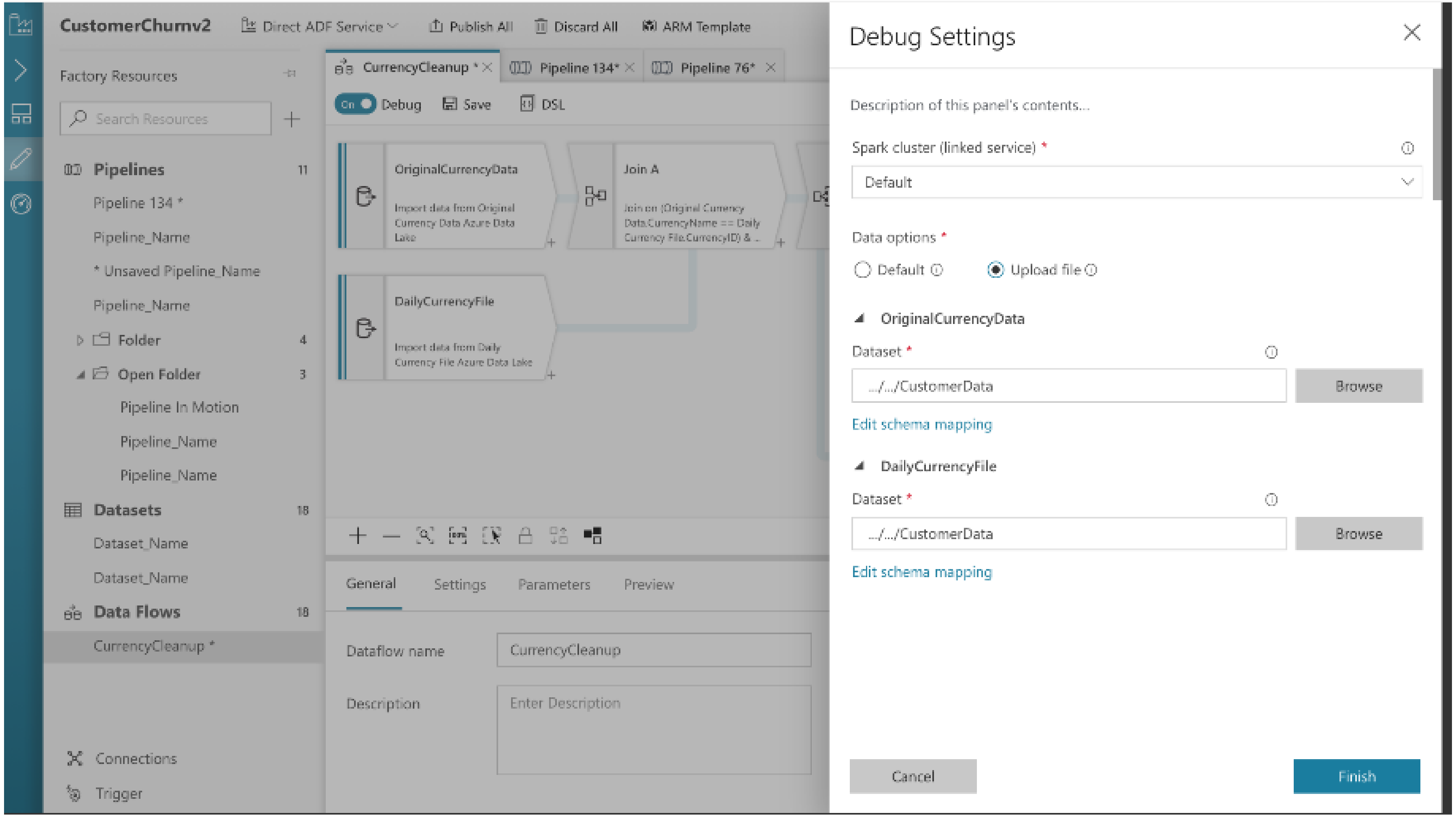


Output stream name * TripData

Source Dataset * trip_data_full

Options ☐ Allow schema drift

Switch to Debug Mode and select sample data to work with for debugging



Dataflow name: CurrencyCleanup

Description: Enter Description

Debug Settings

Description of this panel's contents...

Spark cluster (linked service) *: Default

Data options *: ○ Default ● Upload file

OriginalCurrencyData

Dataset *: .../.../CustomerData

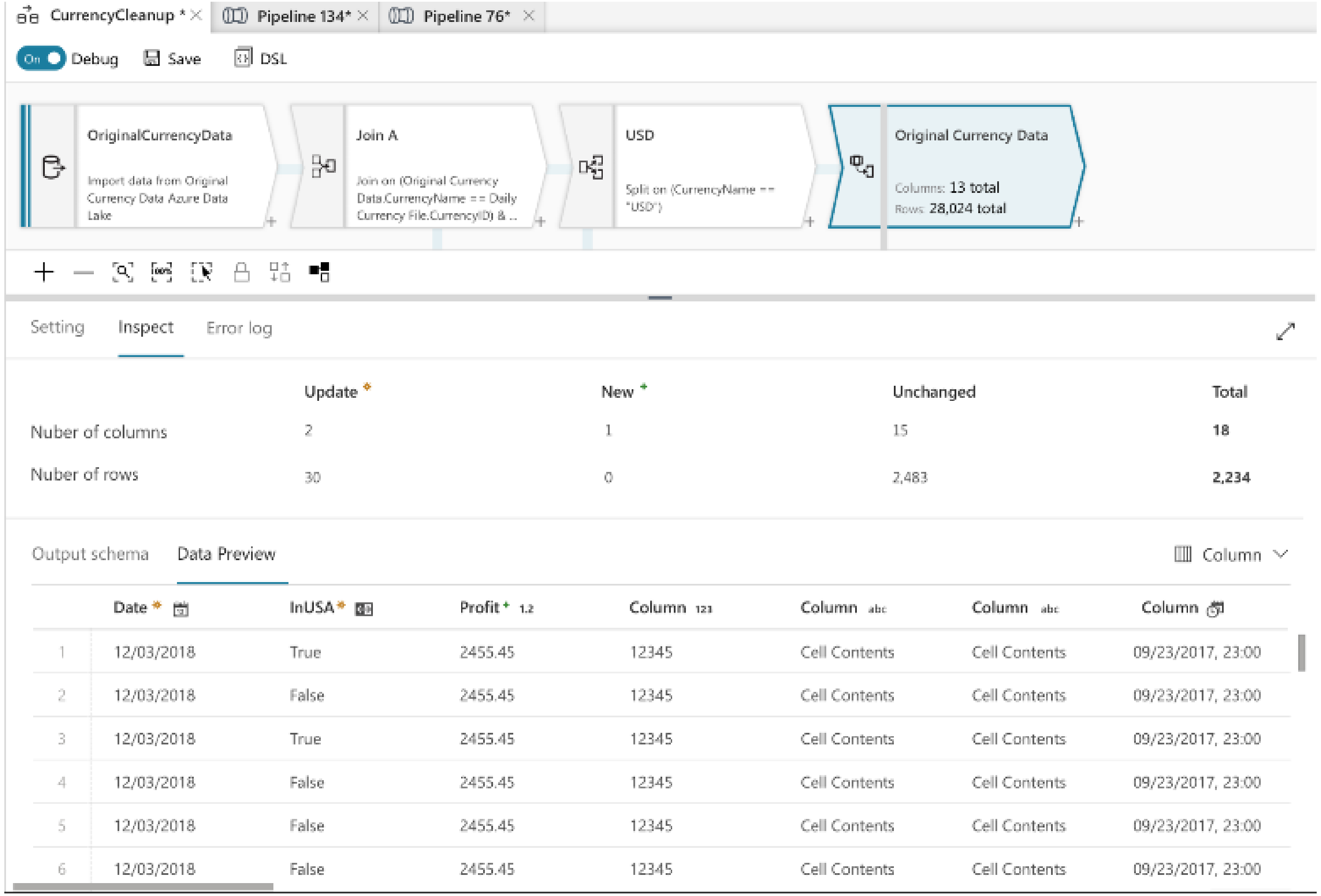
Edit schema mapping

DailyCurrencyFile

Dataset *: .../.../CustomerData

Edit schema mapping

Debug mode provides row-level context and visible results in inspector pane



	Update *	New *	Unchanged	Total
Nuber of columns	2	1	15	**18**
Nuber of rows	30	0	2,483	**2,234**

	Date *	InUSA *	Profit * 1.2	Column 123	Column abc	Column abc	Column
1	12/03/2018	True	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00
2	12/03/2018	False	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00
3	12/03/2018	True	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00
4	12/03/2018	False	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00
5	12/03/2018	False	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00
6	12/03/2018	False	2455.45	12345	Cell Contents	Cell Contents	09/23/2017, 23:00

Questions