DATA WRANGLING FOR ETL ENTHUSIASTS

MOHAMED KABIRUDDIN

CLOUD SOLUTIONS ARCHITECT (DATA)

MICROSOFT

ETL PROCESS

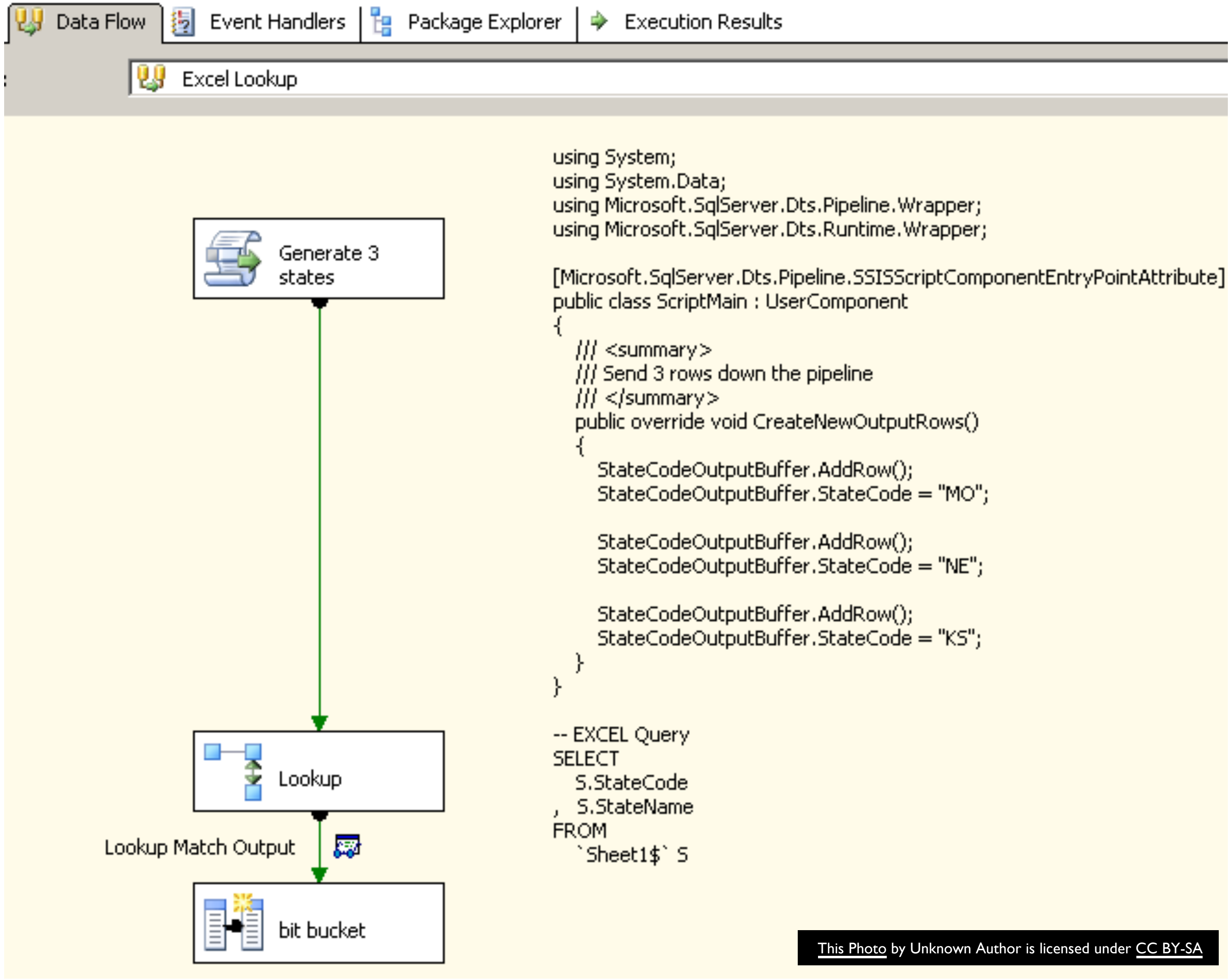
LANDING

STAGING

DIMENSIONAL MODEL

DID I JUST SPEND 10 HOURS PERFECTING THAT LOOKUP?

AND STILL APPLY INDEXES TO GAIN PERFORMANCE?



```
using System;
using System.Data;
using Microsoft.SqlServer.Dts.Pipeline.Wrapper;
using Microsoft.SqlServer.Dts.Runtime.Wrapper;

[Microsoft.SqlServer.Dts.Pipeline.SSISScriptComponentEntryPointAttribute]
public class ScriptMain : UserComponent
{
    /// <summary>
    /// Send 3 rows down the pipeline
    /// </summary>
    public override void CreateNewOutputRows()
    {
        StateCodeOutputBuffer.AddRow();
        StateCodeOutputBuffer.StateCode = "MO";

        StateCodeOutputBuffer.AddRow();
        StateCodeOutputBuffer.StateCode = "NE";

        StateCodeOutputBuffer.AddRow();
        StateCodeOutputBuffer.StateCode = "KS";
    }
}

-- EXCEL Query
SELECT
    S.StateCode
,   S.StateName
FROM
    `Sheet1$` S
```

CONSTRUCT THAT POWERFUL SURROGATE KEY

ULATIMATE GOAL:

KEEP THE
DATA WAREHOUSE

1. UPDATED
2. RELEVANT
3. OPERATIONAL

DATA STAGES IN WRANGLING

RAW

REFINED

PRODUCTION

DATA LAKE DESIGN CONSIDERATIONS

Data Lake Zones

Transient Landing Zone

Temporary storage of data to meet regulatory and quality control requirements. Limited access. May not be required depending on requirements.

Raw Zone

Original source of data ready for consumption. Metadata publicly available but access to data still limited.

Trusted Zone

Standardized and enriched datasets ready for consumption to those with appropriate role-based access. Metadata available to all.

Curated/Refined Zone

Data transformed from Trusted Zone to meet specific business requirements.

Sandbox Zone

Playground for Data Scientists for ad hoc exploratory use cases.

Data Governance Considerations

Security and Compliance

Access Control

Encryption

Row-Level Security

Metadata Management

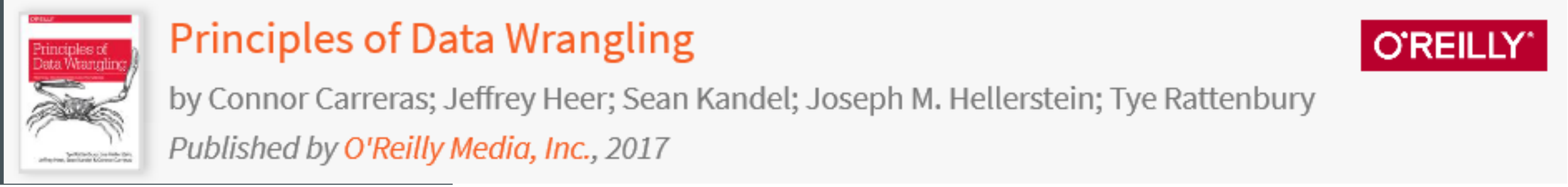
Data Quality

Metadata Management

Lifecycle Management

ARE DATA WRANGLING AND ETL THE SAME THEN?

- Data wrangling is the process involved in transforming or preparing data for analysis
- Consider ETL to be one type of data wrangling, specifically a type of data wrangling managed and overseen by an organization’s shared services or IT organization.
- Data wrangling can also be handled by business users in desktop tools like Excel, or by data scientists in coding languages like Python or R



Principles of Data Wrangling

by Connor Carreras; Jeffrey Heer; Sean Kandel; Joseph M. Hellerstein; Tye Rattenbury

Published by O'Reilly Media, Inc., 2017

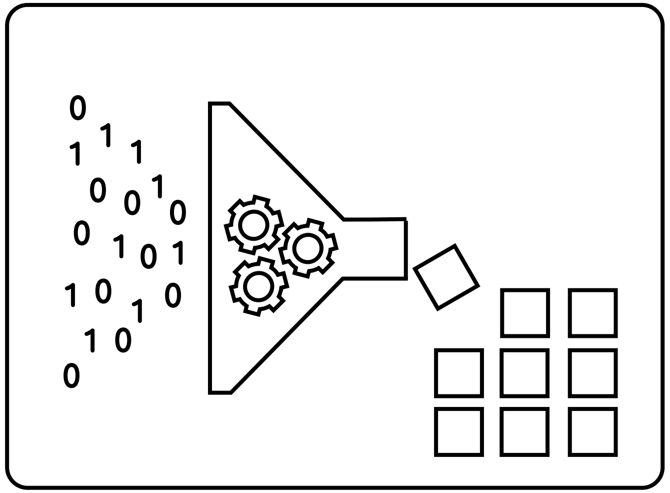
TOOLSET FROM MICROSOFT

- Power BI
- Excel
- SSIS
- T-SQL
- U-SQL
- Polybase
- Azure Data Explorer
- Azure Data Factory
- Azure Stream Analytics
- Azure HD Insight (R & Python)
- Azure Databricks (R, Python and SparkSQL)

MODERN DATA WAREHOUSING

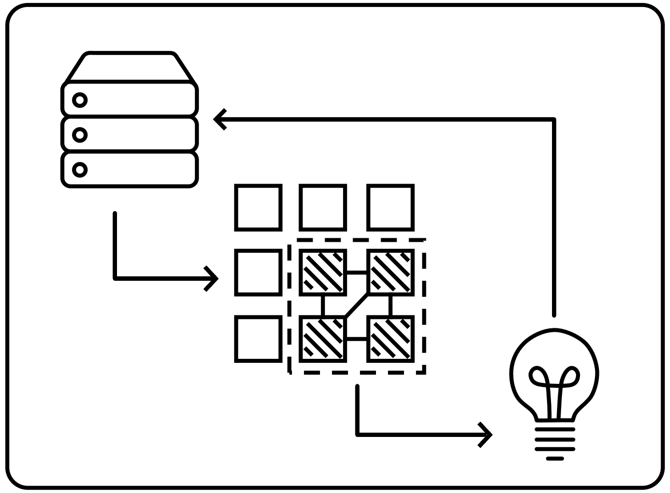
Canonical operations

Load and ingest



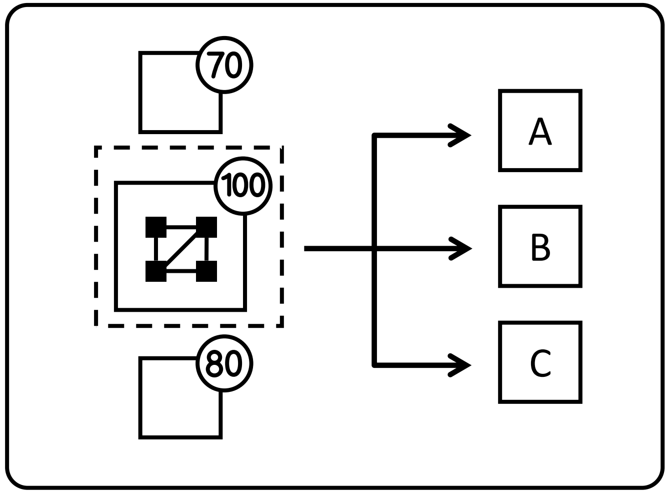
Transfer and store

Process



Process and clean

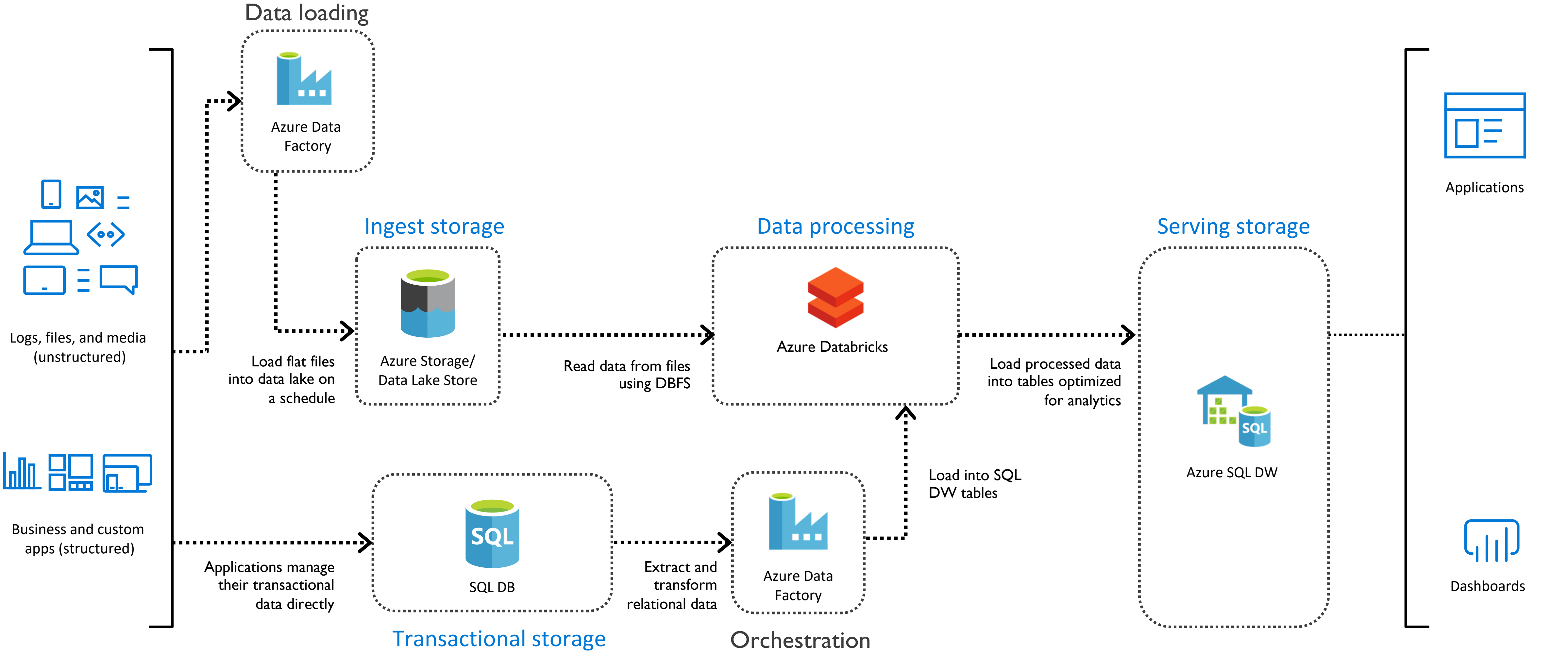
Serve



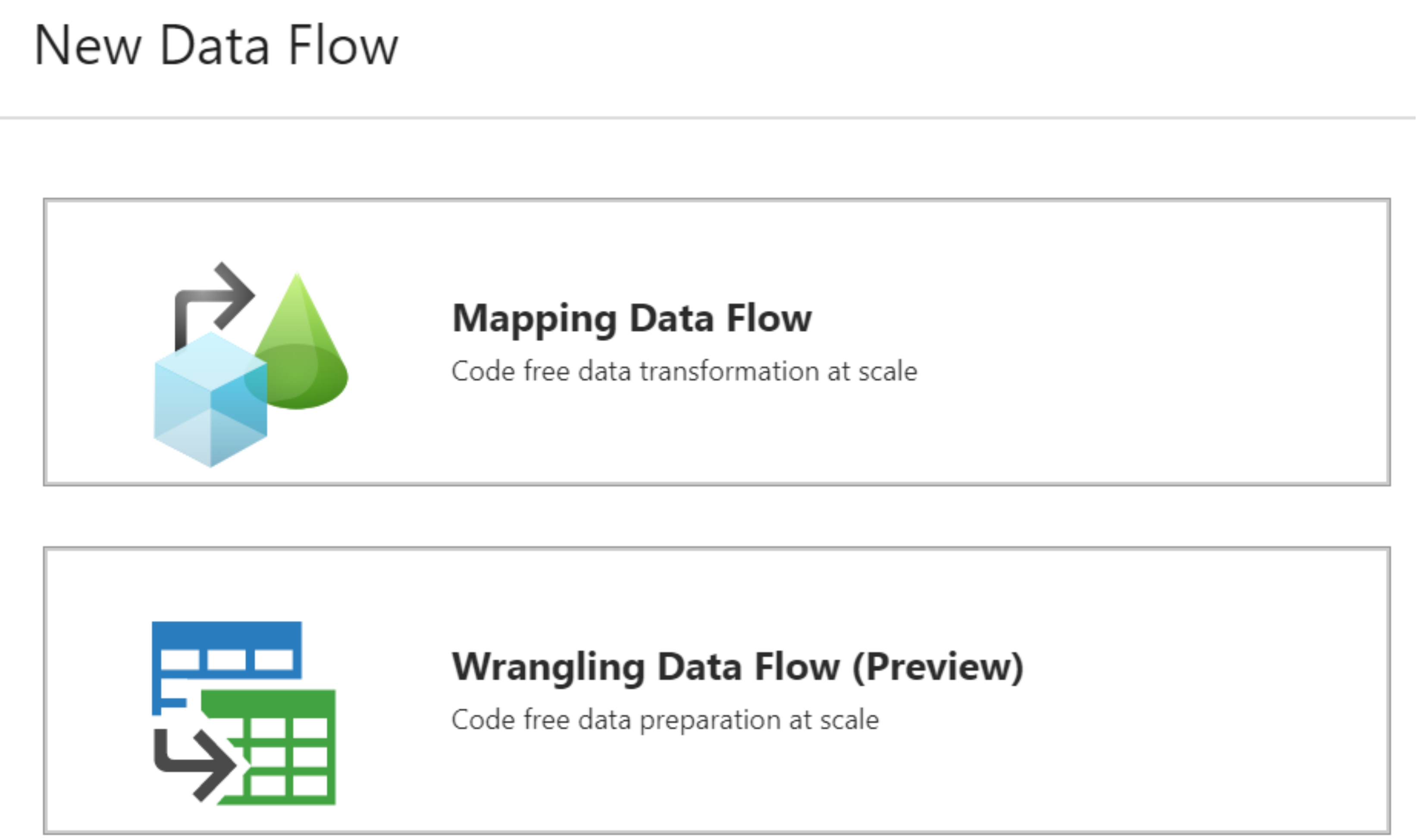
Serve and analyze

MODERN DATA WAREHOUSING PATTERN IN AZURE

Data processing with Azure Databricks


Data loading
Azure Data Factory
Logs, files, and media (unstructured)
Load flat files into data lake on a schedule
Ingest storage
Azure Storage/ Data Lake Store
Read data from files using DBFS
Data processing
Azure Databricks
Load processed data into tables optimized for analytics
Serving storage
Azure SQL DW
SQL
Applications
Business and custom apps (structured)
Applications manage their transactional data directly
SQL
SQL DB
Transactional storage
Extract and transform relational data
Azure Data Factory
Orchestration
Load into SQL DW tables
Dashboards

DATA FACTORY - DATAFLOWS



New Data Flow

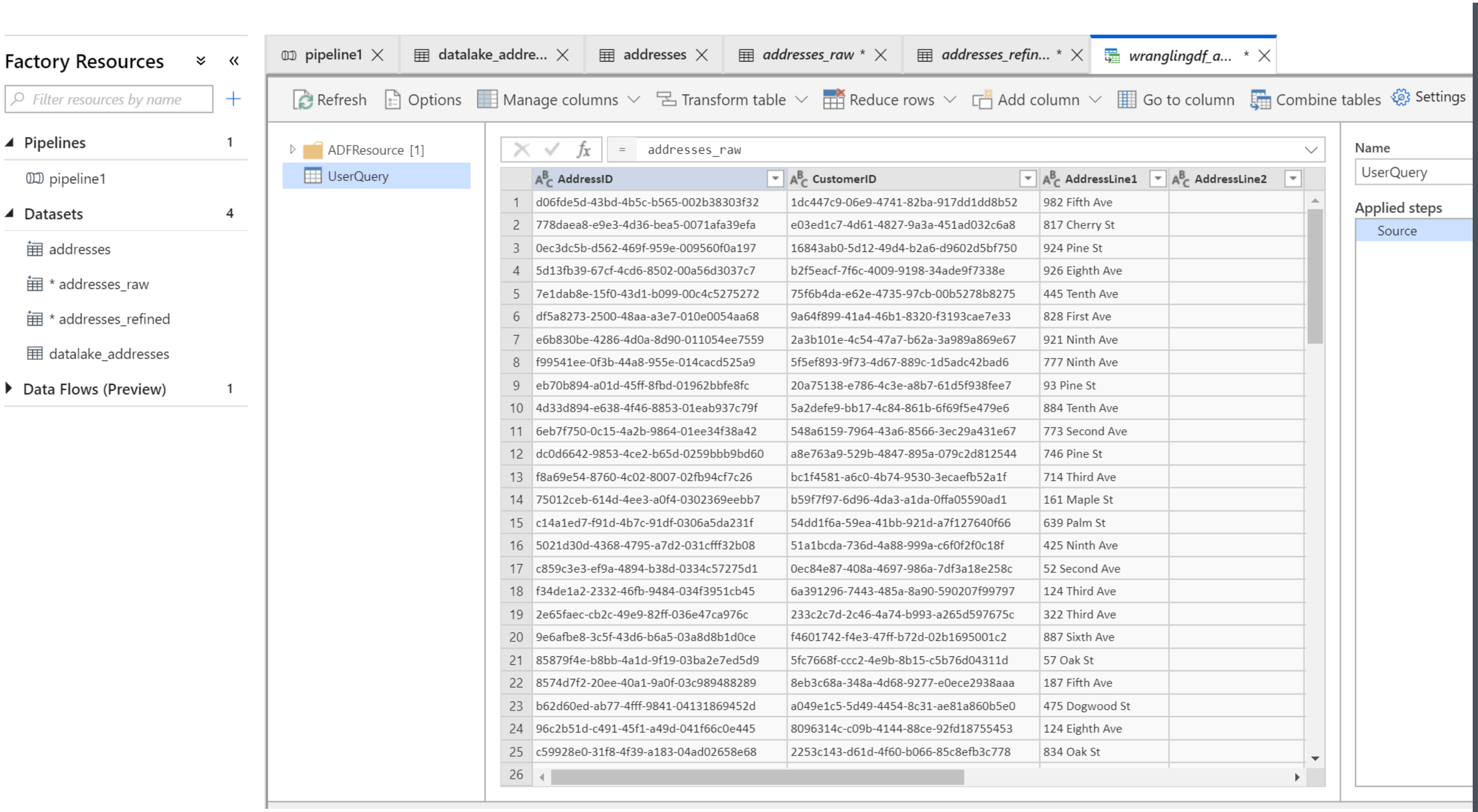
Mapping Data Flow

Code free data transformation at scale

Wrangling Data Flow (Preview)

Code free data preparation at scale

WRANGLING DATAFLOWS



= addresses_raw

	AddressID	CustomerID	AddressLine1	AddressLine2
1	d06fde5d-43bd-4b5c-b565-002b38303f32	1dc447c9-06e9-4741-82ba-917dd1dd8b52	982 Fifth Ave	
2	778daea8-e9e3-4d36-bea5-0071afa39efa	e03ed1c7-4d61-4827-9a3a-451ad032c6a8	817 Cherry St	
3	0ec3dc5b-d562-469f-959e-009560f0a197	16843ab0-5d12-49d4-b2a6-d9602d5bf750	924 Pine St	
4	5d13fb39-67cf-4cd6-8502-00a56d3037c7	b2f5eacf-7f6c-4009-9198-34ade9f7338e	926 Eighth Ave	
5	7e1dab8e-15f0-43d1-b099-00c4c5275272	75f6b4da-e62e-4735-97cb-00b5278b8275	445 Tenth Ave	
6	df5a8273-2500-48aa-a3e7-010e0054aa68	9a64f899-41a4-46b1-8320-f3193cae7e33	828 First Ave	
7	e6b830be-4286-4d0a-8d90-011054ee7559	2a3b101e-4c54-47a7-b62a-3a989a869e67	921 Ninth Ave	
8	f99541ee-0f3b-44a8-955e-014cacd525a9	5f5ef893-9f73-4d67-889c-1d5adc42bad6	777 Ninth Ave	
9	eb70b894-a01d-45ff-8fbd-01962bbfe8fc	20a75138-e786-4c3e-a8b7-61d5f938fee7	93 Pine St	
10	4d33d894-e638-4f46-8853-01eab937c79f	5a2defe9-bb17-4c84-861b-6f69f5e479e6	884 Tenth Ave	
11	6eb7f750-0c15-4a2b-9864-01ee34f38a42	548a6159-7964-43a6-8566-3ec29a431e67	773 Second Ave	
12	dc0d6642-9853-4ce2-b65d-0259bbb9bd60	a8e763a9-529b-4847-895a-079c2d812544	746 Pine St	
13	f8a69e54-8760-4c02-8007-02fb94cf7c26	bc1f4581-a6c0-4b74-9530-3ecaefb52a1f	714 Third Ave	
14	75012ceb-614d-4ee3-a0f4-0302369eebb7	b59f7f97-6d96-4da3-a1da-0ffa05590ad1	161 Maple St	
15	c14a1ed7-f91d-4b7c-91df-0306a5da231f	54dd1f6a-59ea-41bb-921d-a7f127640f66	639 Palm St	
16	5021d30d-4368-4795-a7d2-031cfff32b08	51a1bcda-736d-4a88-999a-c6f0f2f0c18f	425 Ninth Ave	
17	c859c3e3-ef9a-4894-b38d-0334c57275d1	0ec84e87-408a-4697-986a-7df3a18e258c	52 Second Ave	
18	f34de1a2-2332-46fb-9484-034f3951cb45	6a391296-7443-485a-8a90-590207f99797	124 Third Ave	
19	2e65faec-cb2c-49e9-82ff-036e47ca976c	233c2c7d-2c46-4a74-b993-a265d597675c	322 Third Ave	
20	9e6afbe8-3c5f-43d6-b6a5-03a8d8b1d0ce	f4601742-f4e3-47ff-b72d-02b1695001c2	887 Sixth Ave	
21	85879f4e-b8bb-4a1d-9f19-03ba2e7ed5d9	5fc7668f-ccc2-4e9b-8b15-c5b76d04311d	57 Oak St	
22	8574d7f2-20ee-40a1-9a0f-03c989488289	8eb3c68a-348a-4d68-9277-e0ece2938aaa	187 Fifth Ave	
23	b62d60ed-ab77-4fff-9841-04131869452d	a049e1c5-5d49-4454-8c31-ae81a860b5e0	475 Dogwood St	
24	96c2b51d-c491-45f1-a49d-041f66c0e445	8096314c-c09b-4144-88ce-92fd18755453	124 Eighth Ave	
25	c59928e0-31f8-4f39-a183-04ad02658e68	2253c143-d61d-4f60-b066-85c8efb3c778	834 Oak St	
26				

Name
UserQuery

Applied steps
Source

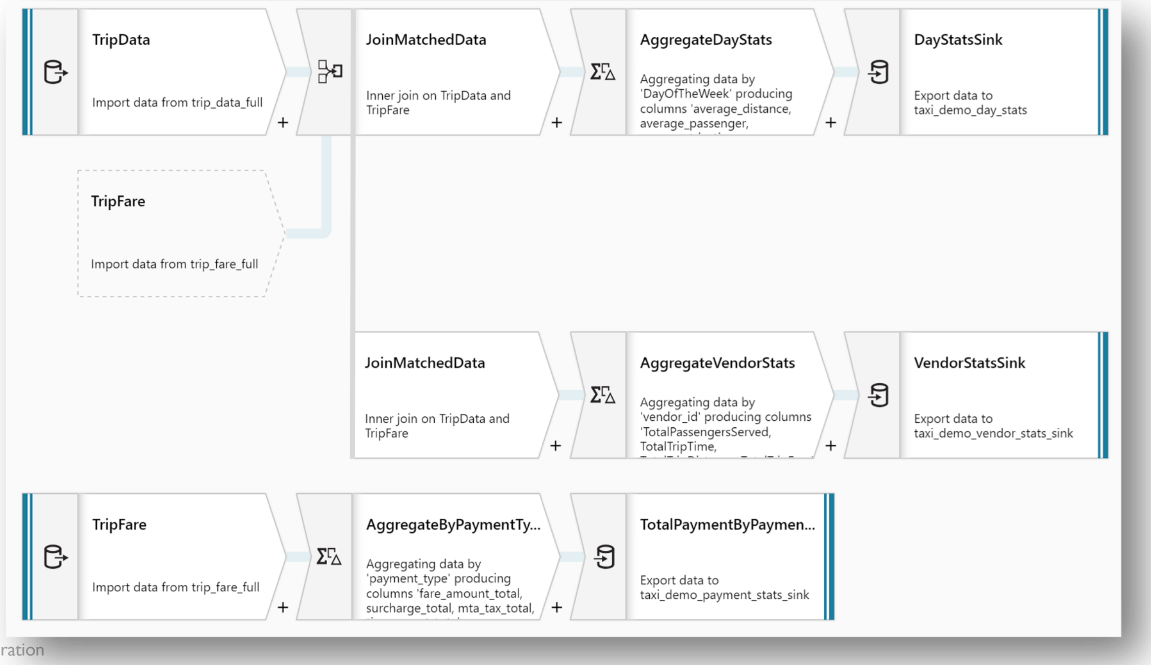
MAPPING DATA FLOW

No-code data transformation @ scale

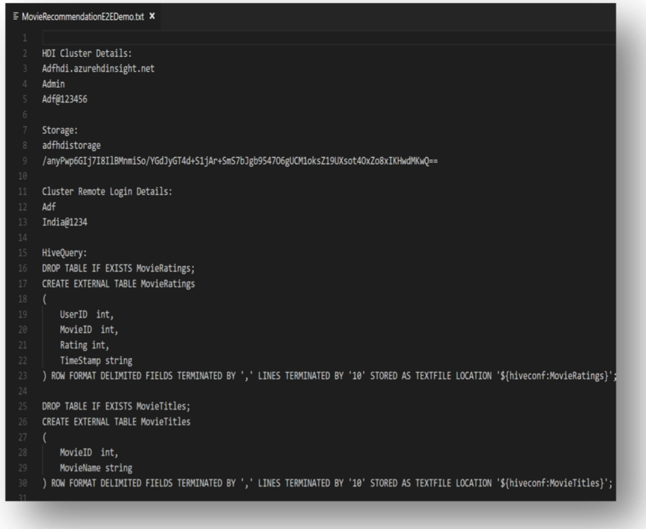
Data cleansing, transformation, aggregation, conversion, etc.

Cloud scale via Spark execution

Easily build resilient data flows


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

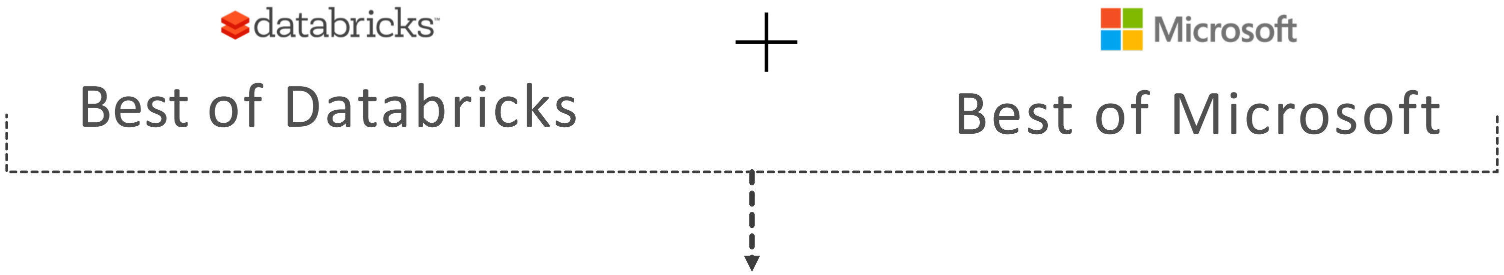
... not


MovieRecommendationE2EDemo.txt

```

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

Storage:
adfhdistorage
/anyPwp6GIj7I8IlBMnmiSo/YGdJyGT4d+S1jAr+SmS7bJgb954706gUCM1oksZ19UXsot4OxZo8xIKHwdMKwQ==

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 DATABRICKS

A fast, easy and collaborative Apache® Spark™ based analytics platform optimized for Azure



databricks

Best of Databricks

+

Microsoft

Best of Microsoft

- Designed in collaboration with the founders of Apache Spark
- One-click set up; streamlined workflows
- Interactive workspace that enables collaboration between data scientists, data engineers, and business analysts.
- Native integration with Azure services (Power BI, SQL DW, Cosmos DB, ADLS, Azure Storage, Azure Data Factory, Azure AD, Event Hub, IoT Hub, HDInsight Kafka, SQL DB)
- Enterprise-grade Azure security (Active Directory integration, compliance, enterprise-grade SLAs)

AZURE DATABRICKS NOTEBOOKS

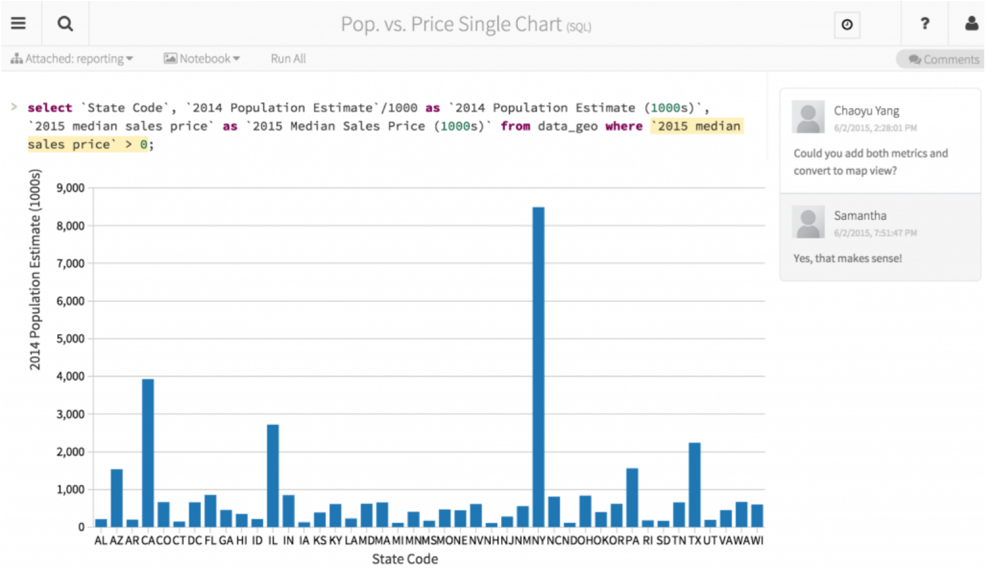
Notebooks are a popular way to develop, and run, Spark Applications

Notebooks are not only for authoring Spark applications but can be *run/executed directly* on clusters

- Shift+Enter
- click the ▶ at the top right of the cell in a notebook
- Submit via Job

Fine grained permissions support so they can be *securely shared* with colleagues for collaboration

Notebooks are well-suited for prototyping, rapid development, exploration, discovery and iterative development


Pop. vs. Price Single Chart (SQL)
?
Attached: reporting
Notebook
Run All
Comments

```
select `State Code`, `2014 Population Estimate`/1000 as `2014 Population Estimate (1000s)`,
`2015 median sales price` as `2015 Median Sales Price (1000s)` from data_geo where `2015 median
sales price` > 0;
```

2014 Population Estimate (1000s)
9,000
8,000
7,000
6,000
5,000
4,000
3,000
2,000
1,000
0
AL AZ AR CA CO CT DC FL GA HI ID IL IN IA KS KY LA MD MA MI MN MS MO NE NV NH NJ NM NY NC ND OH OK OR PA RI SD TN TX UT VA WA WI
State Code

Chaoyu Yang
6/2/2015, 2:28:01 PM
Could you add both metrics and convert to map view?

Samantha
6/2/2015, 7:51:47 PM
Yes, that makes sense!

With Azure Databricks notebooks you have a default language but you can mix multiple languages in the same notebook:

`%python`	Allows you to execute python code in a notebook (even if that notebook is not python)
`%sql`	Allows you to execute sql code in a notebook (even if that notebook is not sql).
`%r`	Allows you to execute r code in a notebook (even if that notebook is not r).
`%scala`	Allows you to execute scala code in a notebook (even if that notebook is not scala).
`%sh`	Allows you to execute shell code in your notebook.
`%fs`	Allows you to use Databricks Utilities - dbutils filesystem commands.
`%md`	To include rendered markdown

Microsoft