Microsoft Ignite
Australia 2017

Orchestrating Big Data Pipelines with Azure Data Factory

DA332
Lace Lofranco

Survey

Microsoft

Session Objective

Learn to design, build and manage big data orchestration pipelines using Azure Data Factory

Agenda

Design

Big data pipelines
Lamda Architecture
Data Factory Concepts

Build

Data Movement
Data Transformation

Manage

Monitor pipeline health
Developer tools

Compare

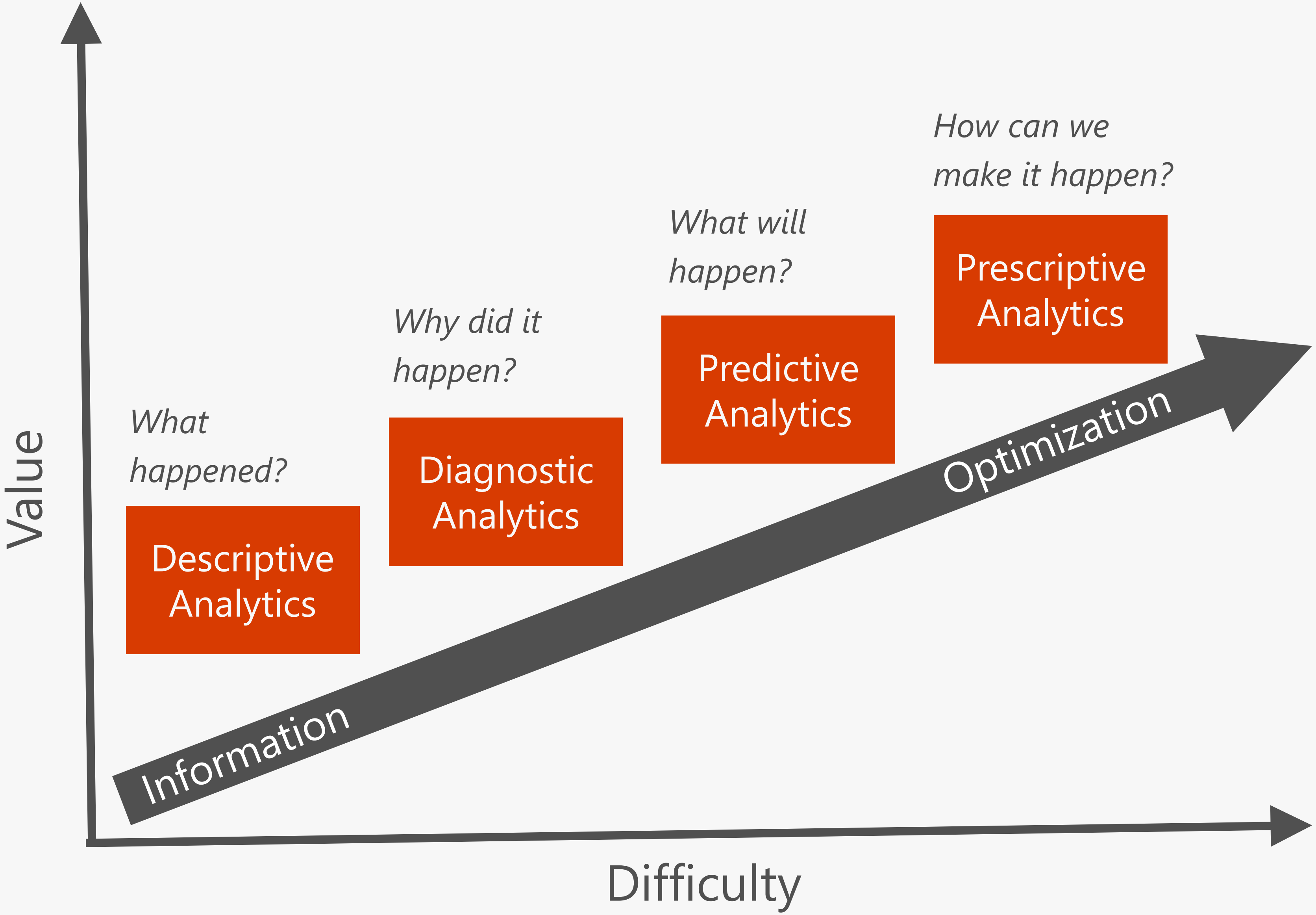
Data Factory vs Oozie

Microsoft Ignite
Australia 2017

Design: Big Data Pipelines

Microsoft

Types of analytics



What happened?

Descriptive Analytics

Why did it happen?

Diagnostic Analytics

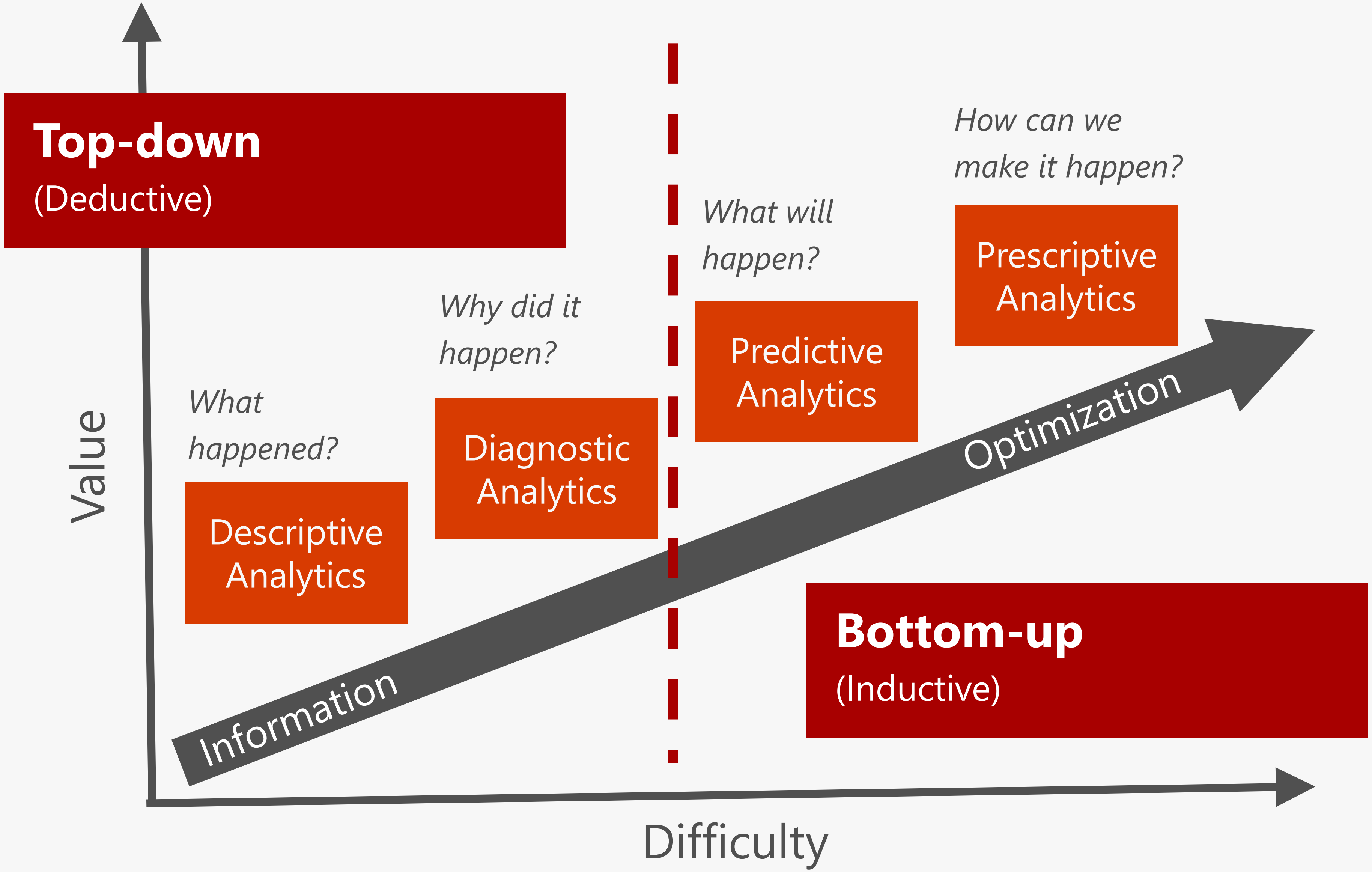
What will happen?

Predictive Analytics

How can we make it happen?

Prescriptive Analytics

Types of analytics



Top-down
(Deductive)

What happened?

Descriptive Analytics

Why did it happen?

Diagnostic Analytics

What will happen?

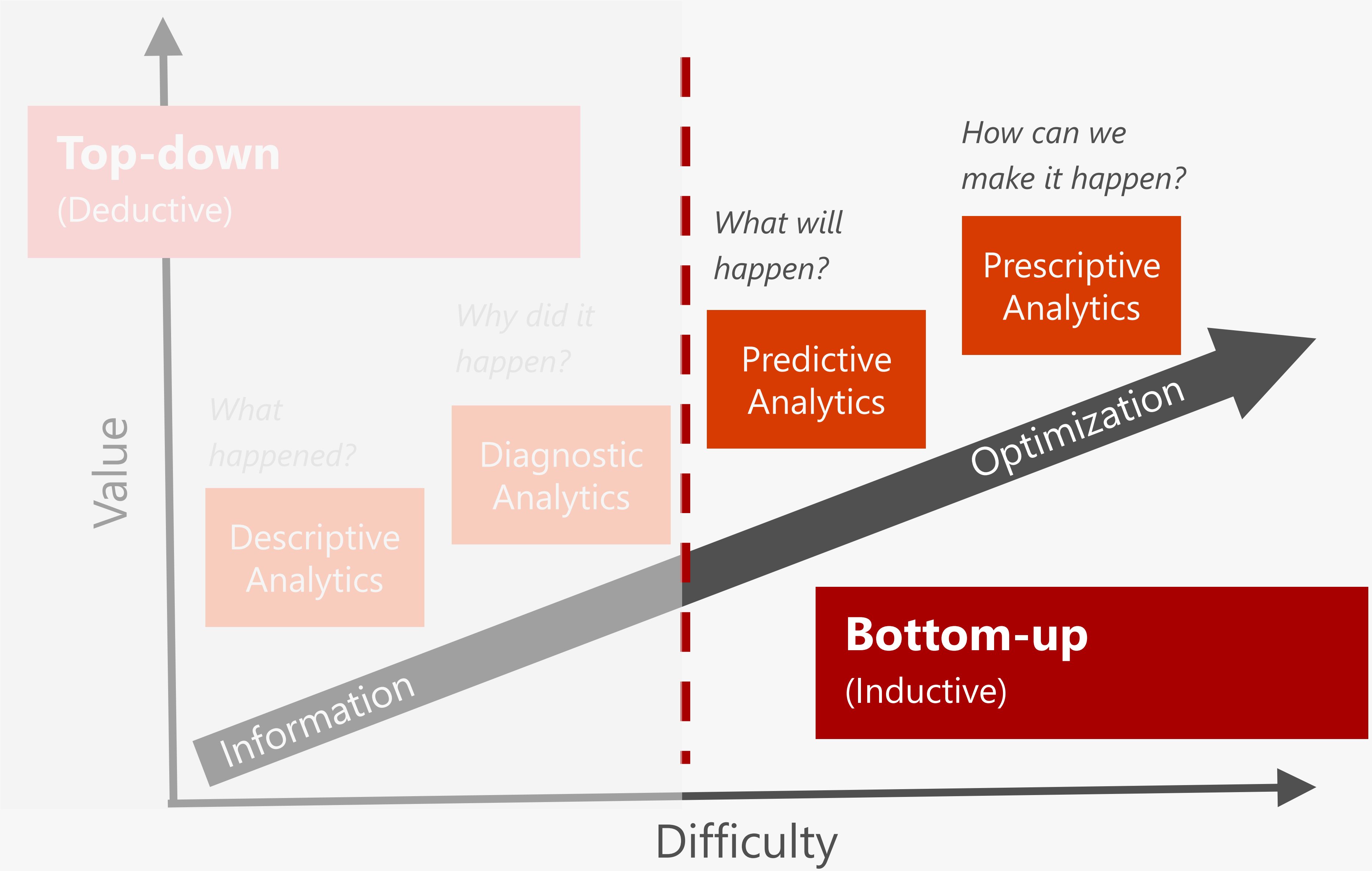
Predictive Analytics

How can we make it happen?

Prescriptive Analytics

Bottom-up
(Inductive)

Types of analytics

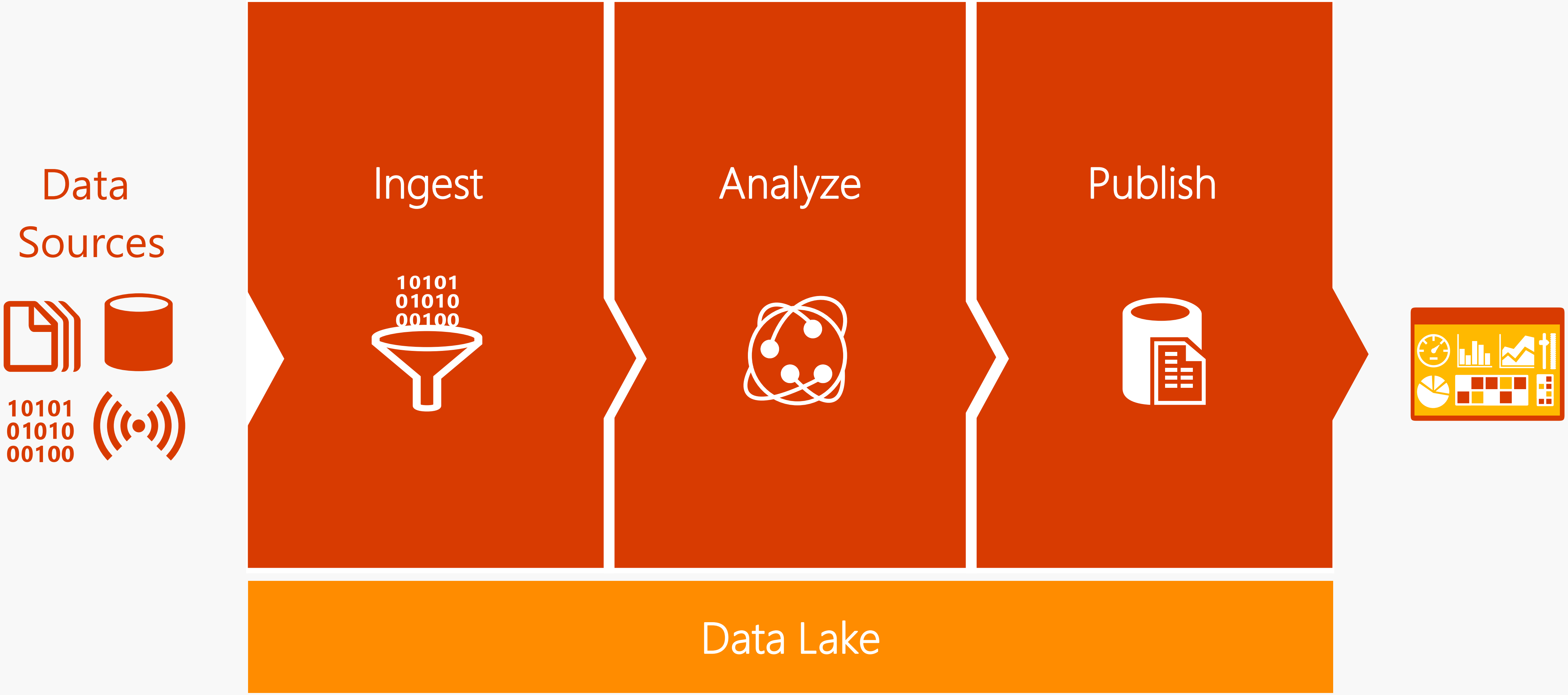


Big Data Pipelines Examples

Velocity, Variety, Volume

- Optimizing Ad revenue
- Demand Forecasting
- Predictive Maintenance
- Supply chain optimization
- Portfolio optimization

Big Data Pipelines



Data Sources

Ingest

Analyze

Publish

Data Lake

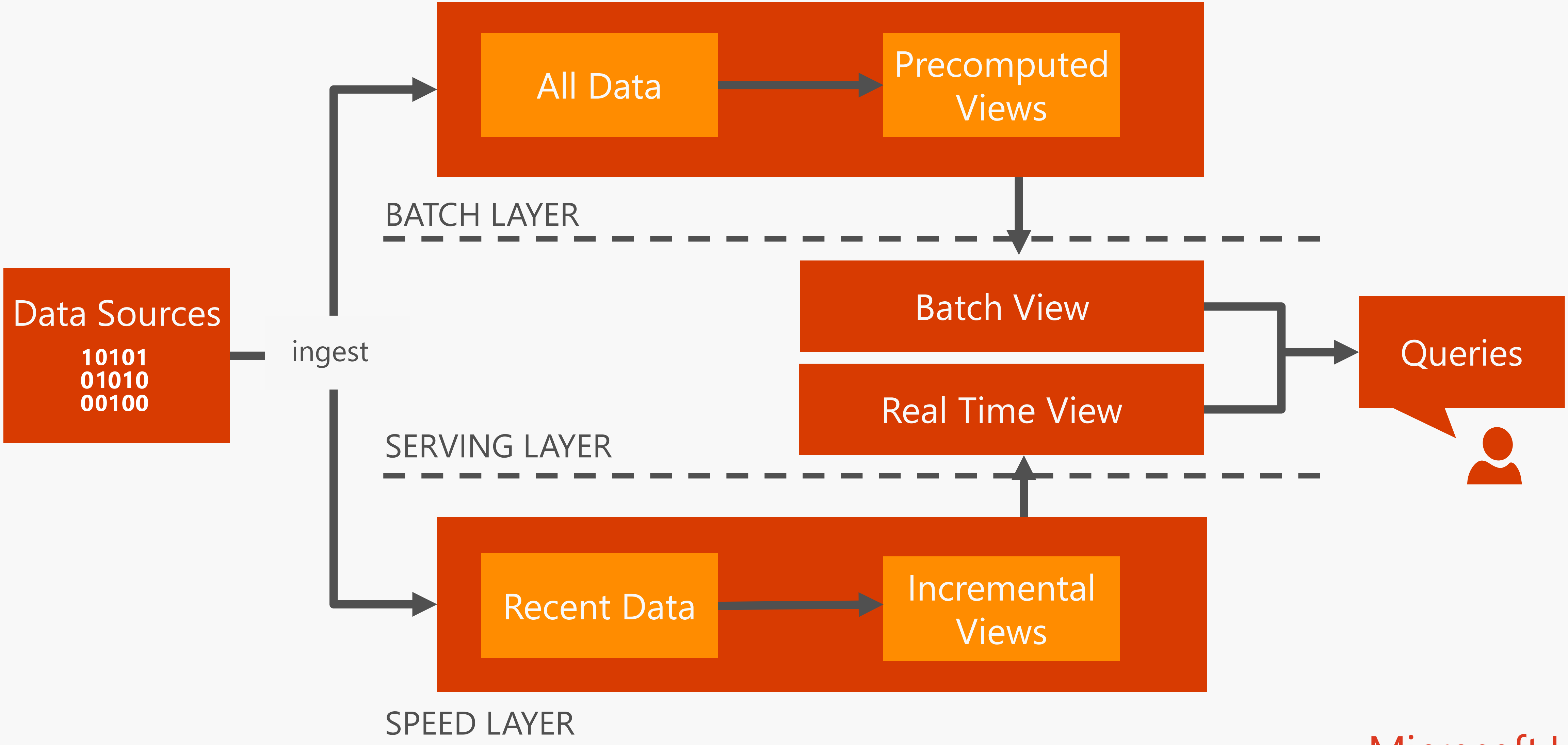
Lamda Architecture

Data architecture for designing Big Data applications

Three layers: Batch, Speed, Serving

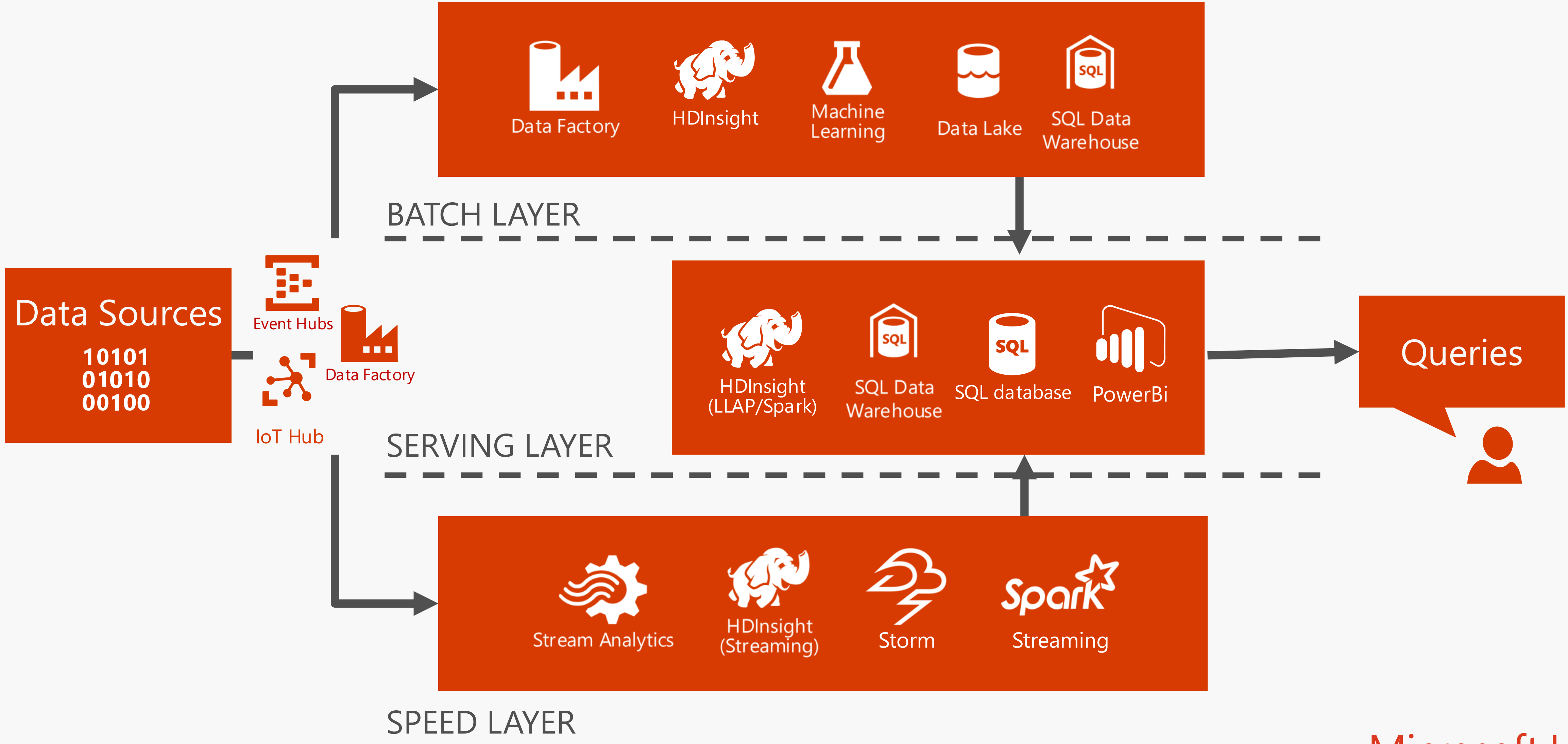
Popularized by Nathan Marz

Lamda Architecture



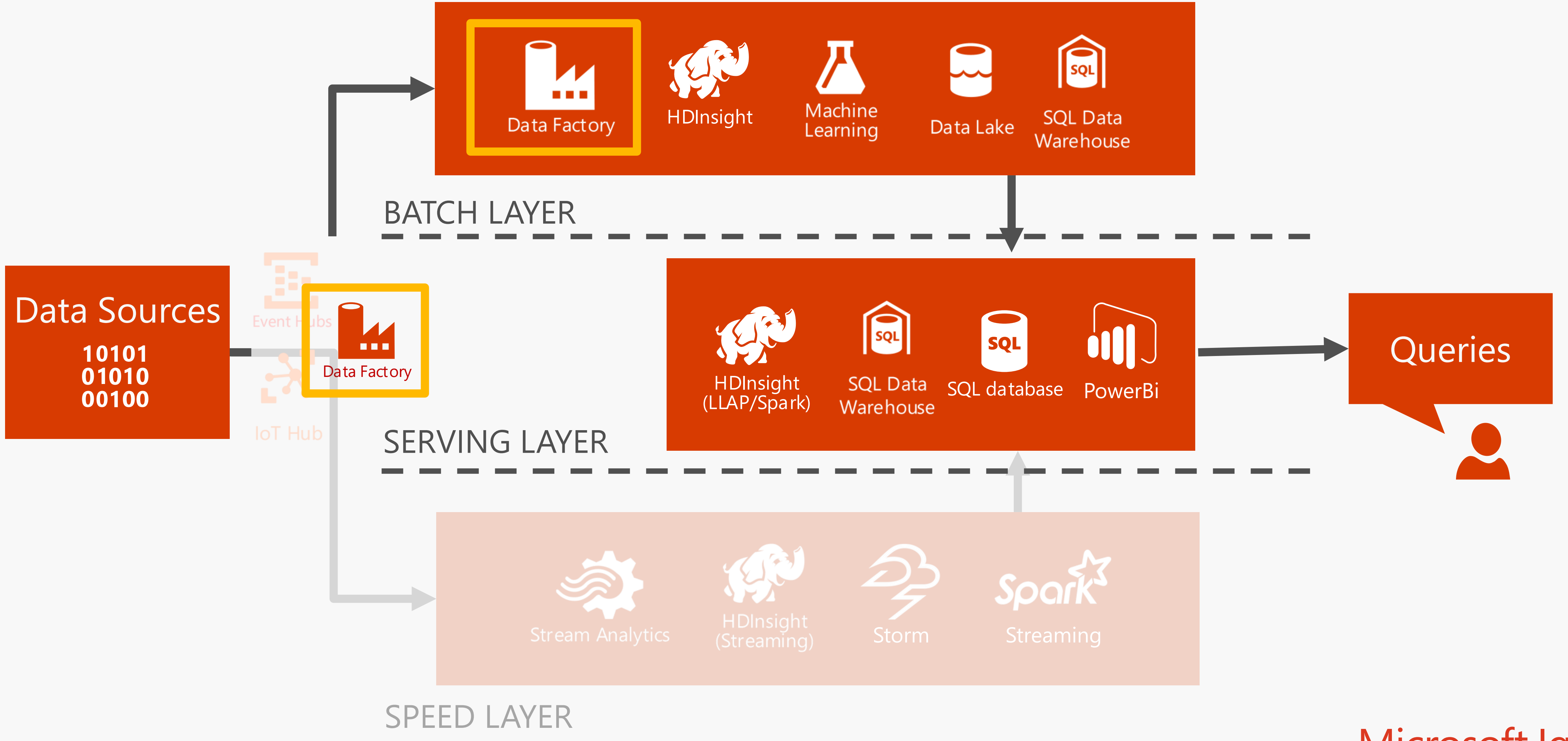
Microsoft Ignite

Lamda Architecture



Microsoft Ignite

Lamda Architecture



Microsoft Ignite

Melbourne Foot Traffic

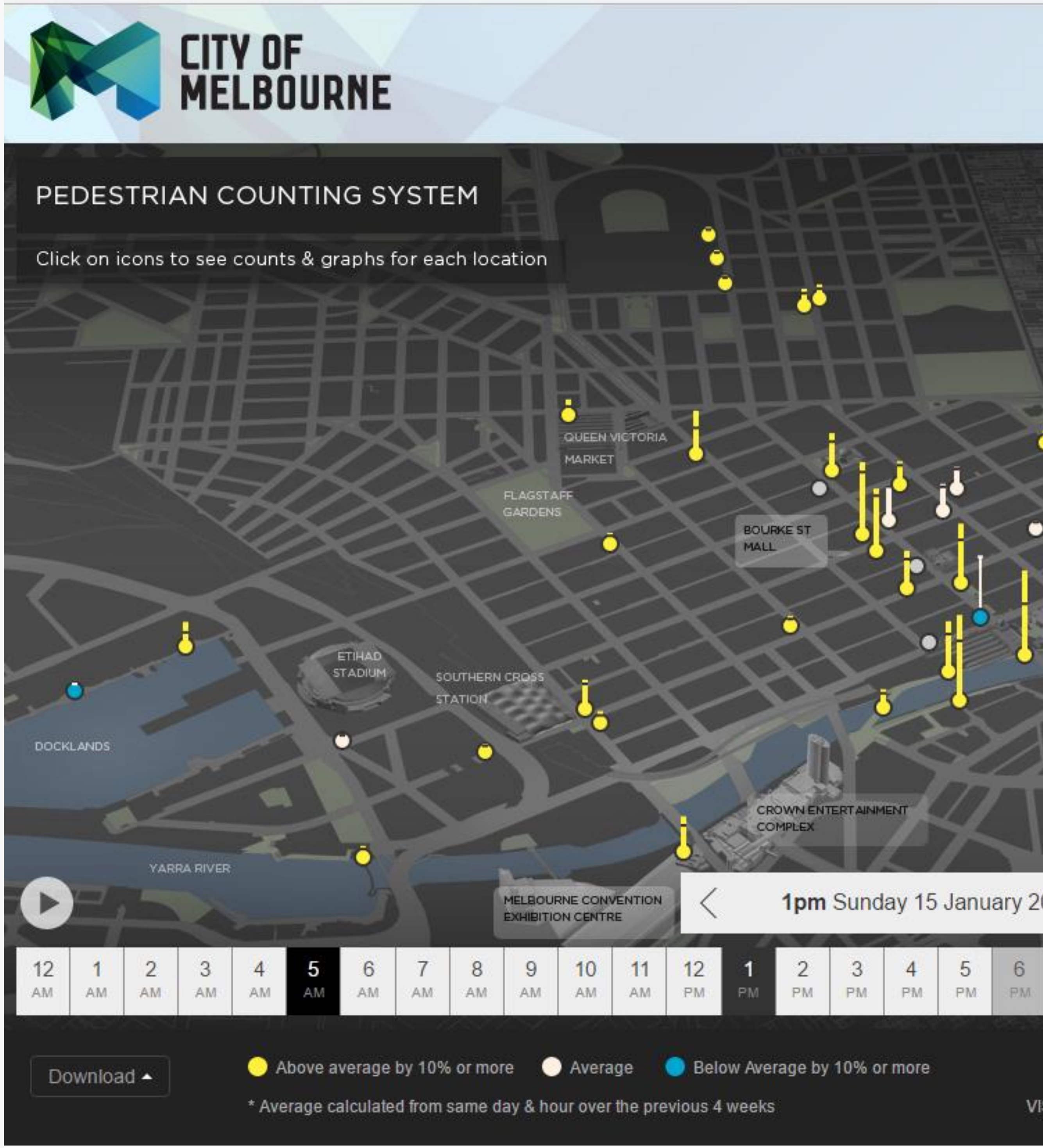
Predicting **tram load** based on **foot traffic**

Model training data from yearly tram load survey conducted by Public Transport Victoria

Data sources:

https://data.melbourne.vic.gov.au

https://www.ptv.vic.gov.au



CITY OF MELBOURNE

PEDESTRIAN COUNTING SYSTEM

Click on icons to see counts & graphs for each location

1pm Sunday 15 January 2

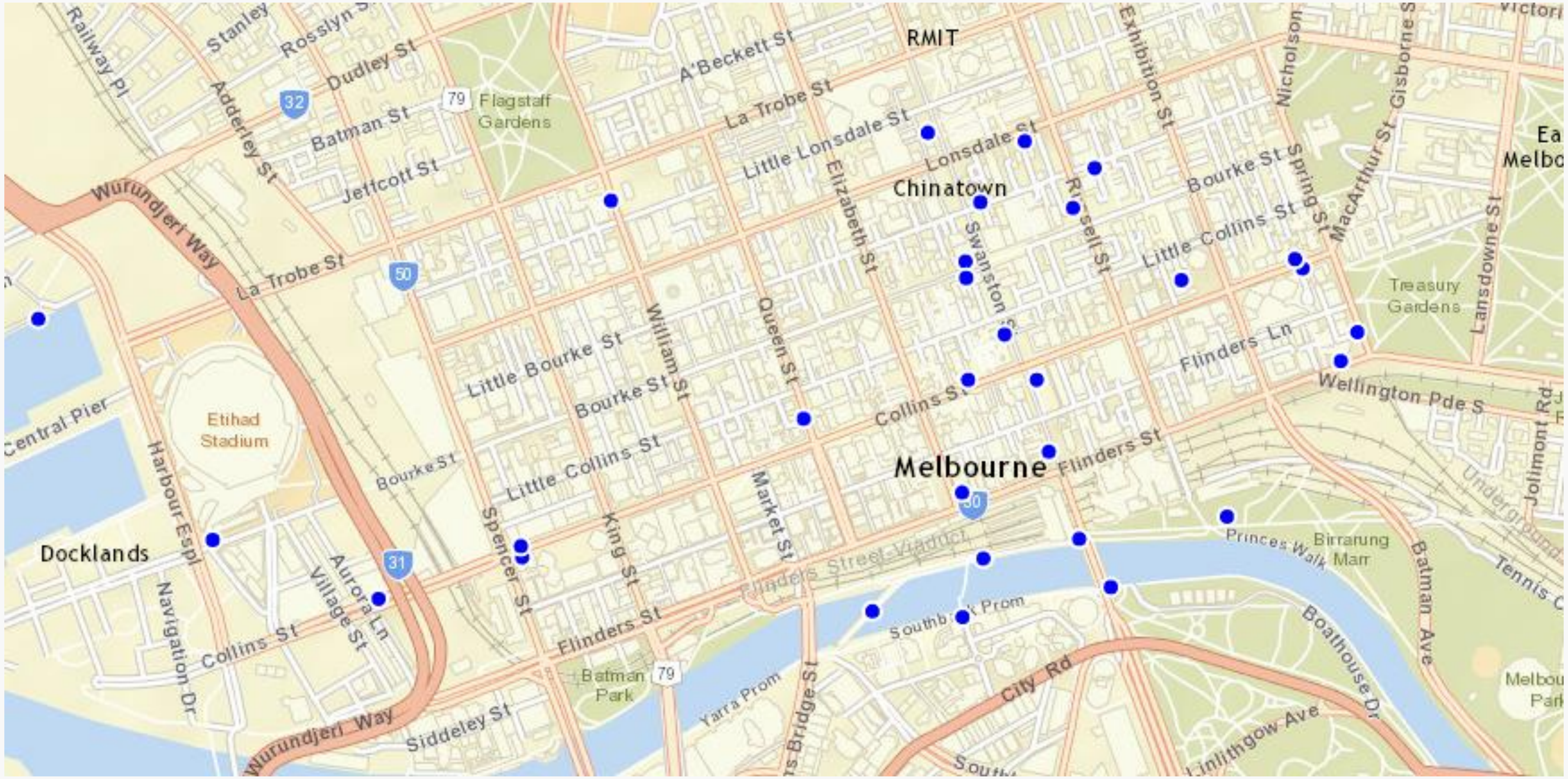
12 AM | 1 AM | 2 AM | 3 AM | 4 AM | 5 AM | 6 AM | 7 AM | 8 AM | 9 AM | 10 AM | 11 AM | 12 PM | 1 PM | 2 PM | 3 PM | 4 PM | 5 PM | 6 PM

Download ▴

● Above average by 10% or more ● Average ● Below Average by 10% or more

* Average calculated from same day & hour over the previous 4 weeks

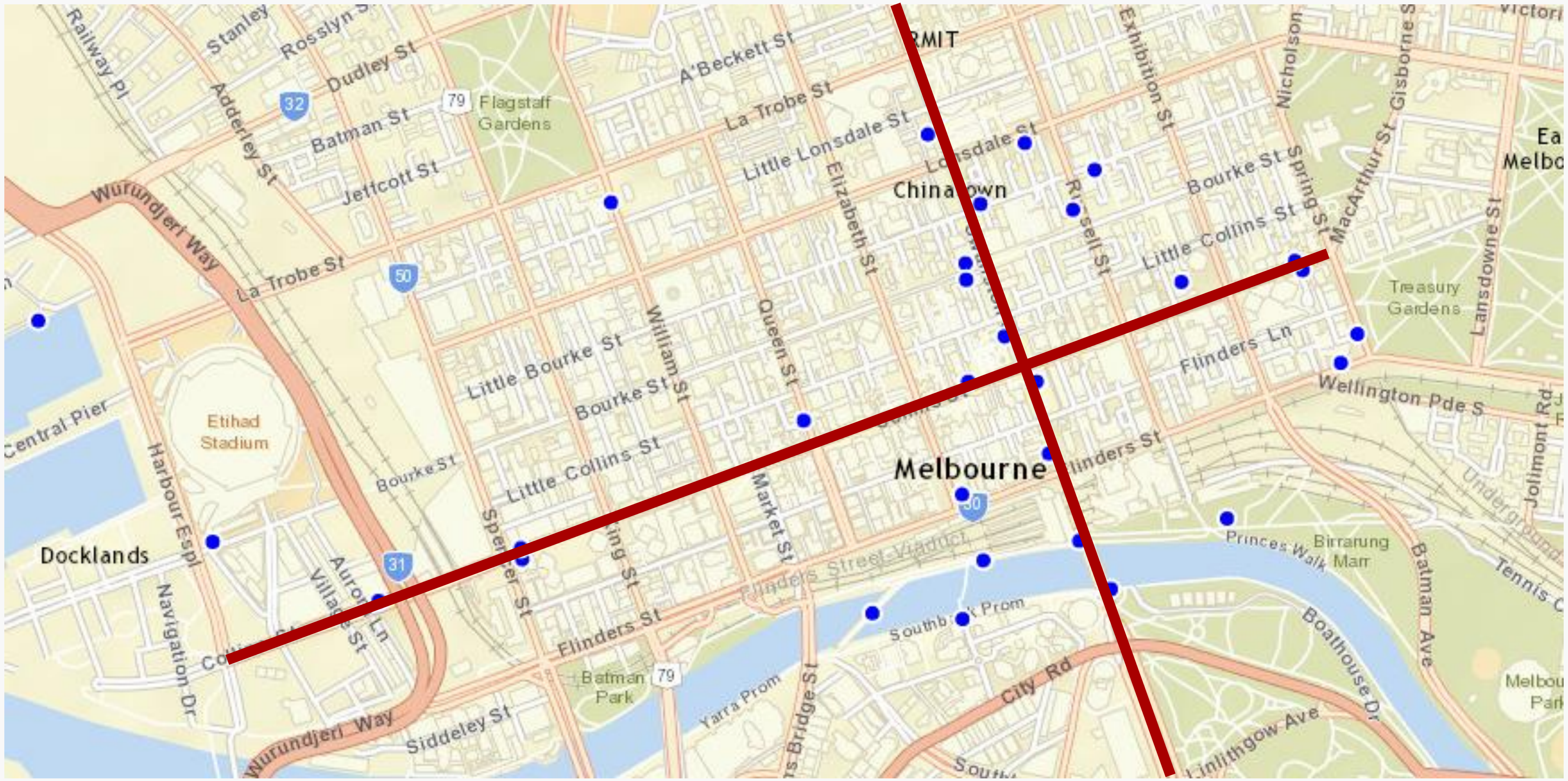
Sensor locations



Source: https://data.melbourne.vic.gov.au

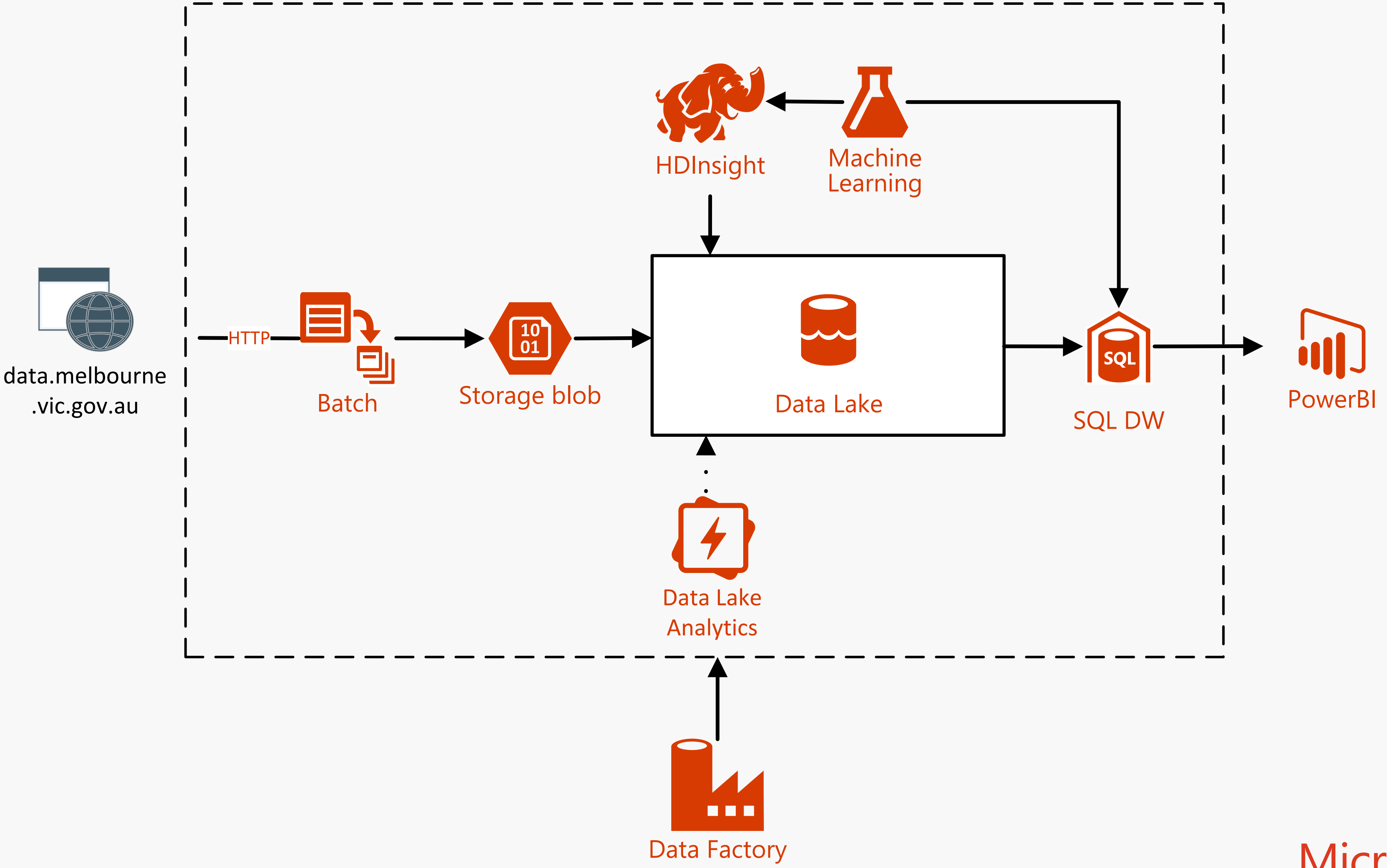
Microsoft Ignite

Sensor locations



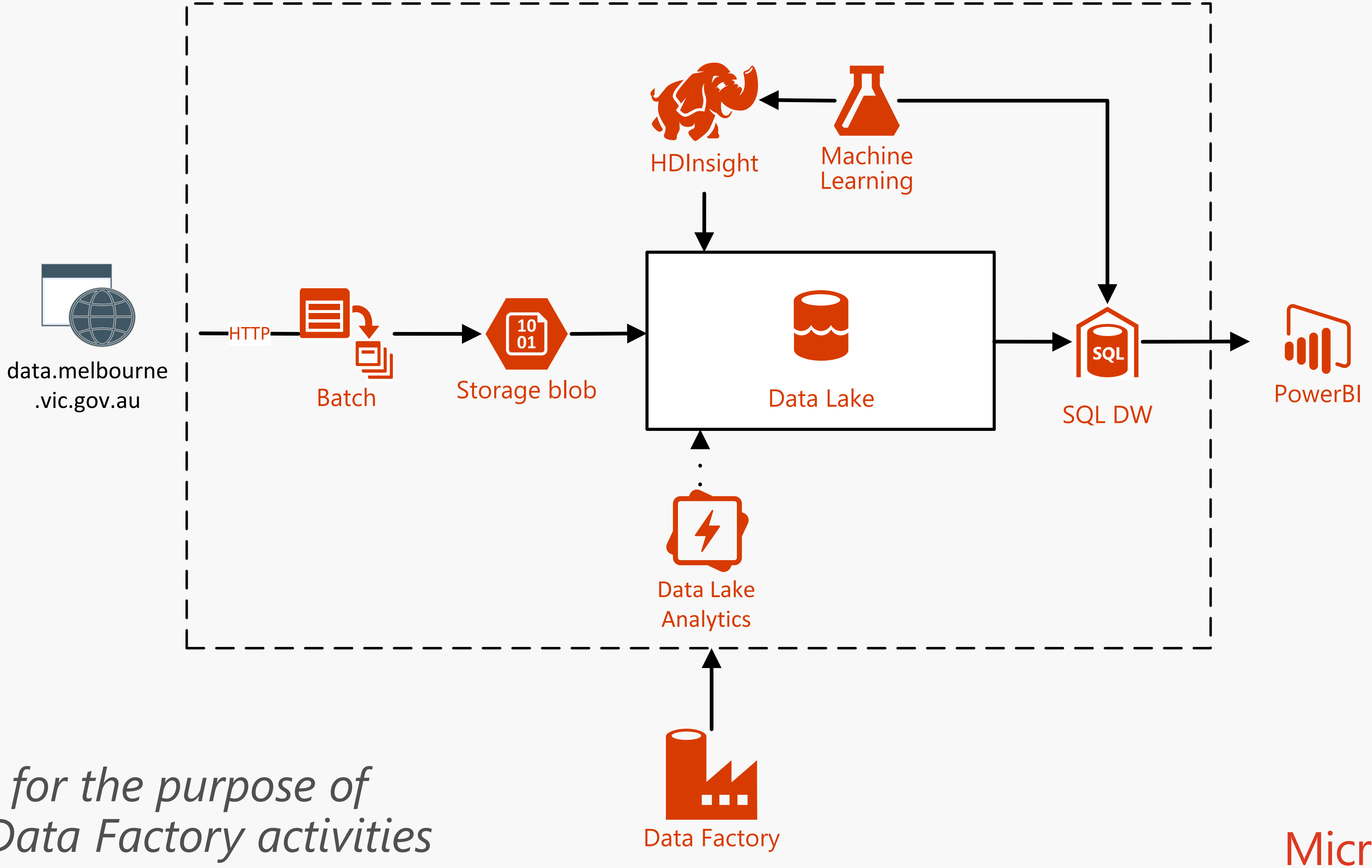
Source: https://data.melbourne.vic.gov.au

Demo Solution Architecture



Microsoft Ignite

Demo Solution Architecture*



**Designed for the purpose of demoing Data Factory activities*

Microsoft Ignite
Australia 2017

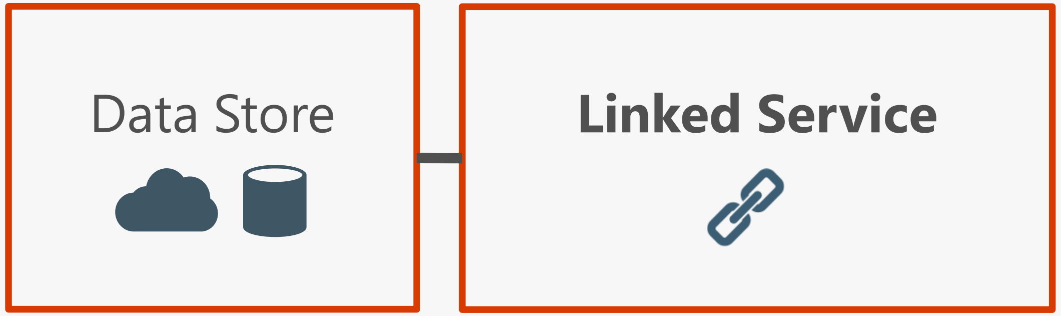
Design: Data Factory

Microsoft

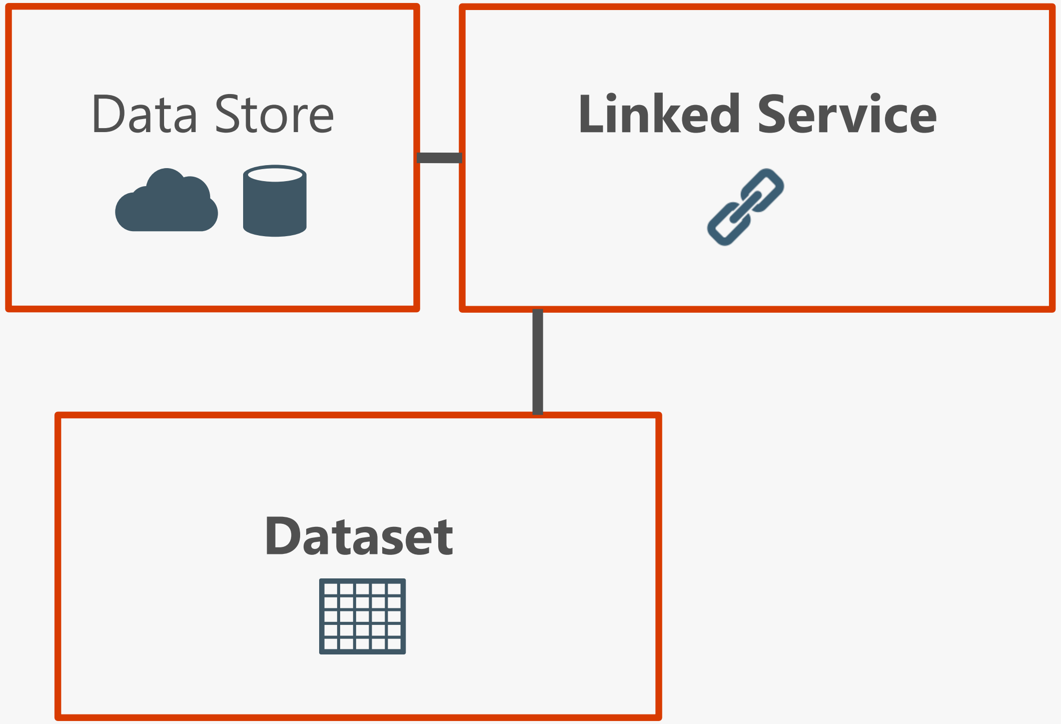
Data Factory Concepts

Data Store

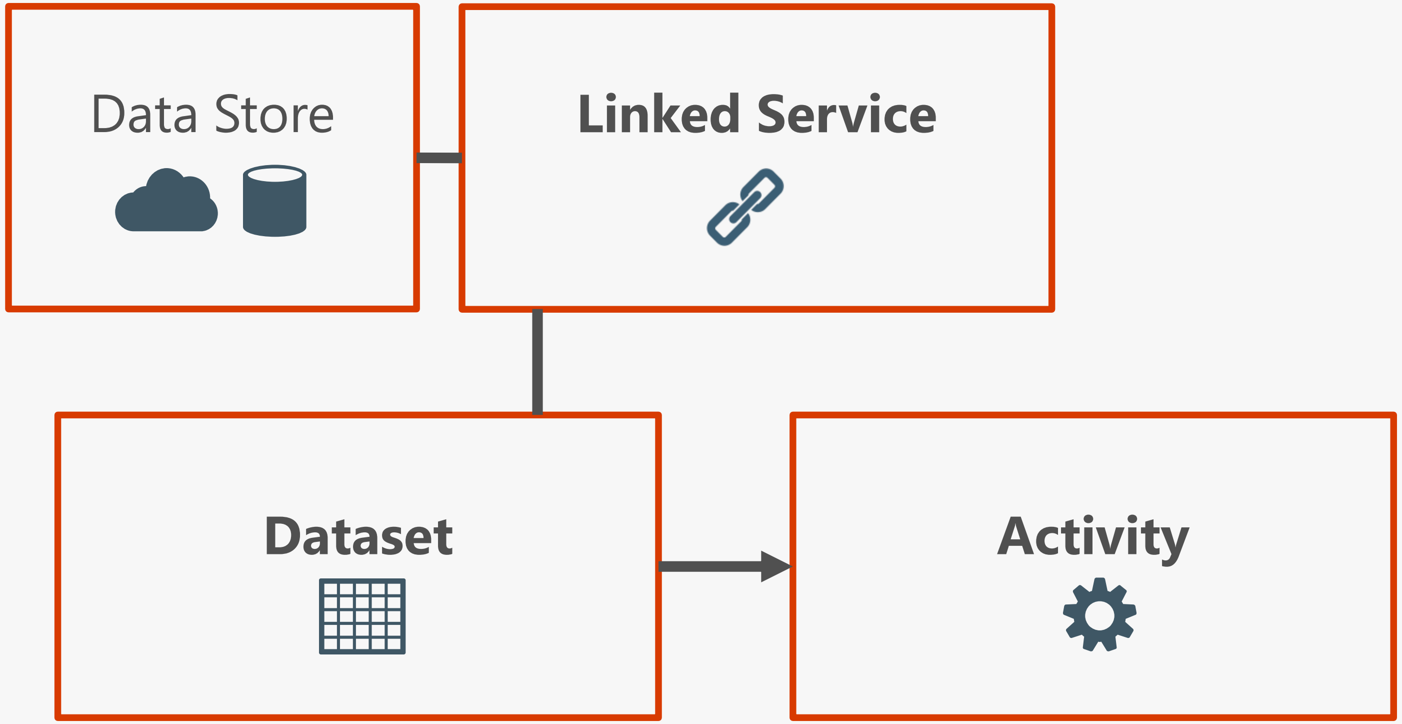
Data Factory Concepts



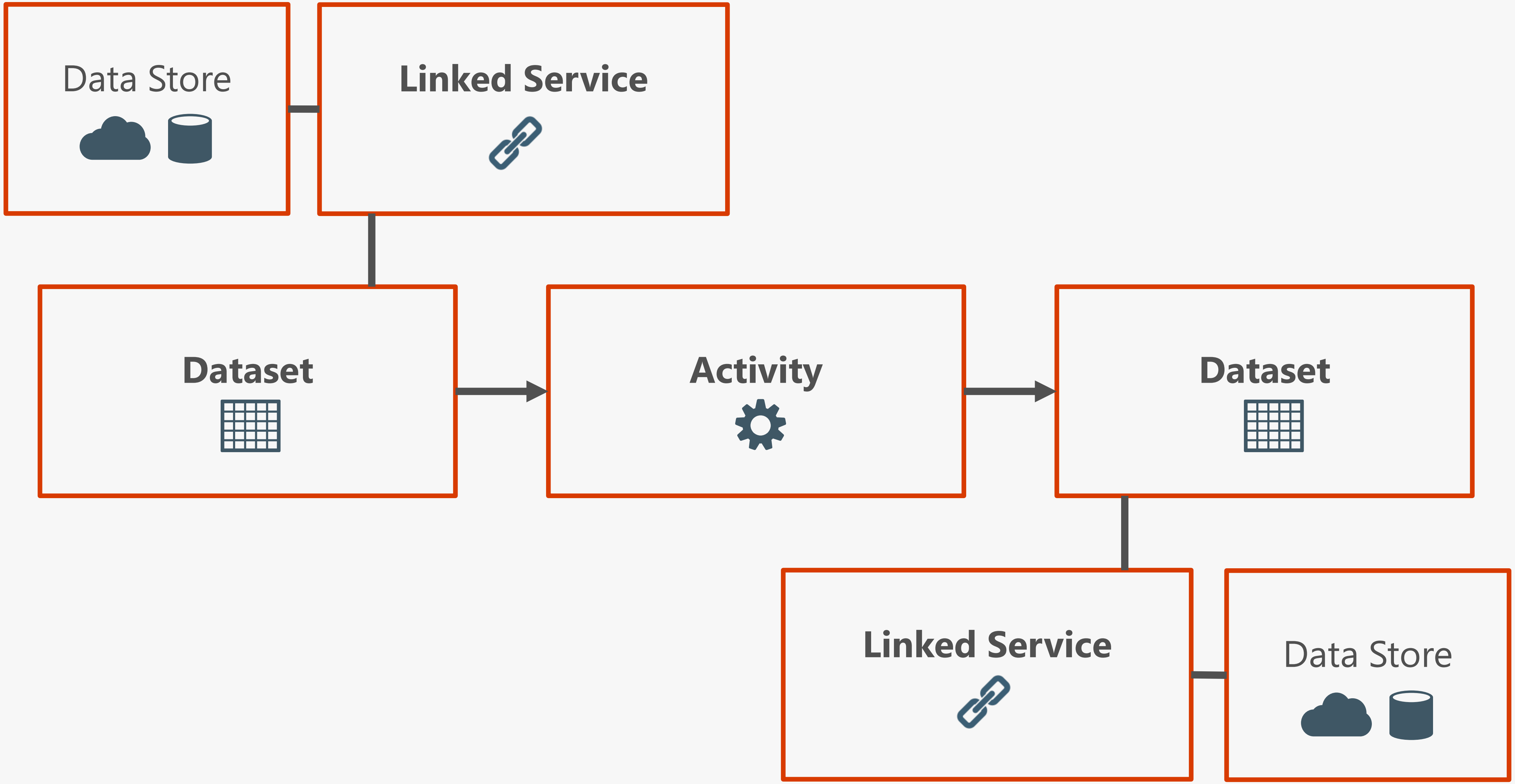
Data Factory Concepts



Data Factory Concepts



Data Factory Concepts



Data Store

Linked Service

Dataset

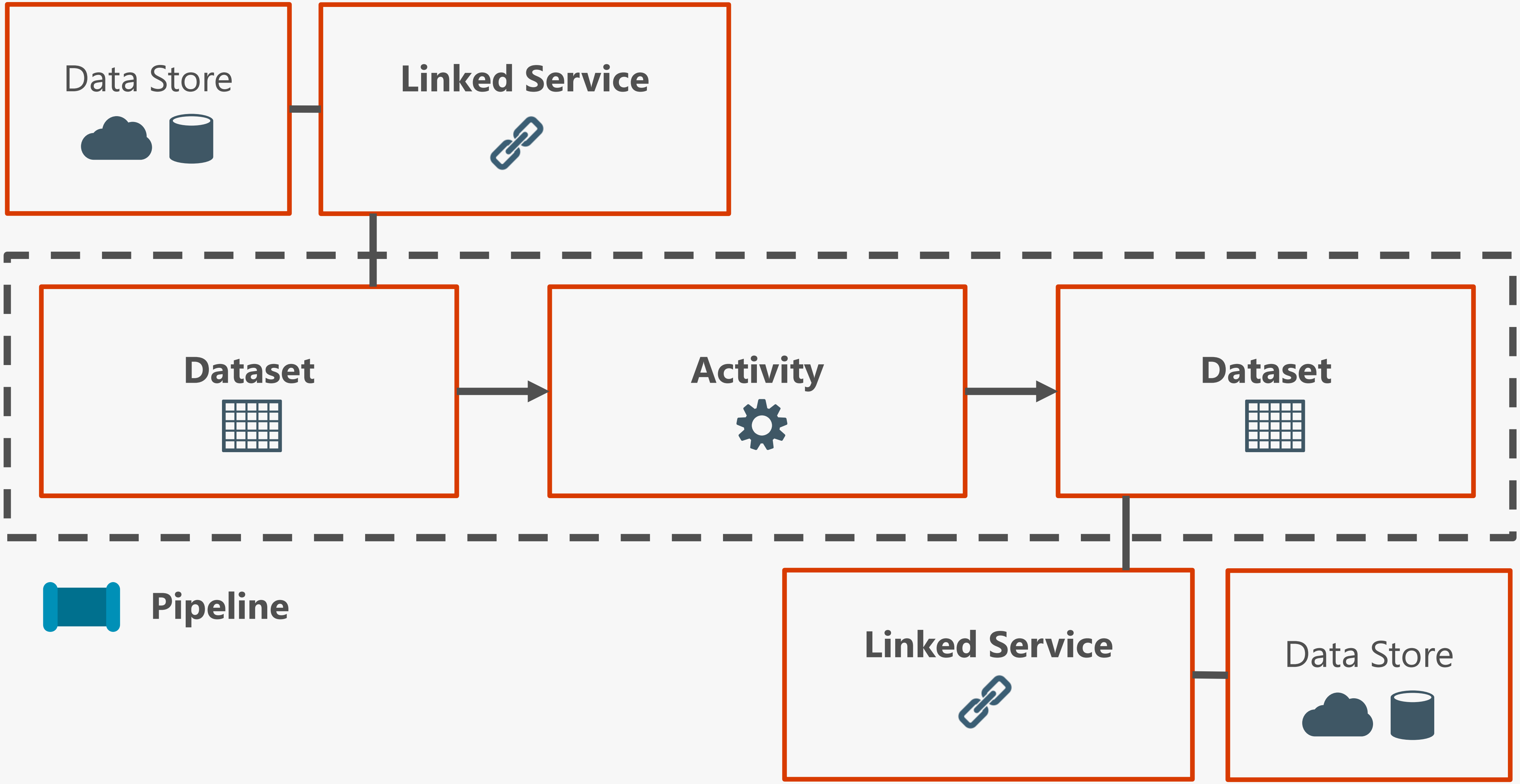
Activity

Dataset

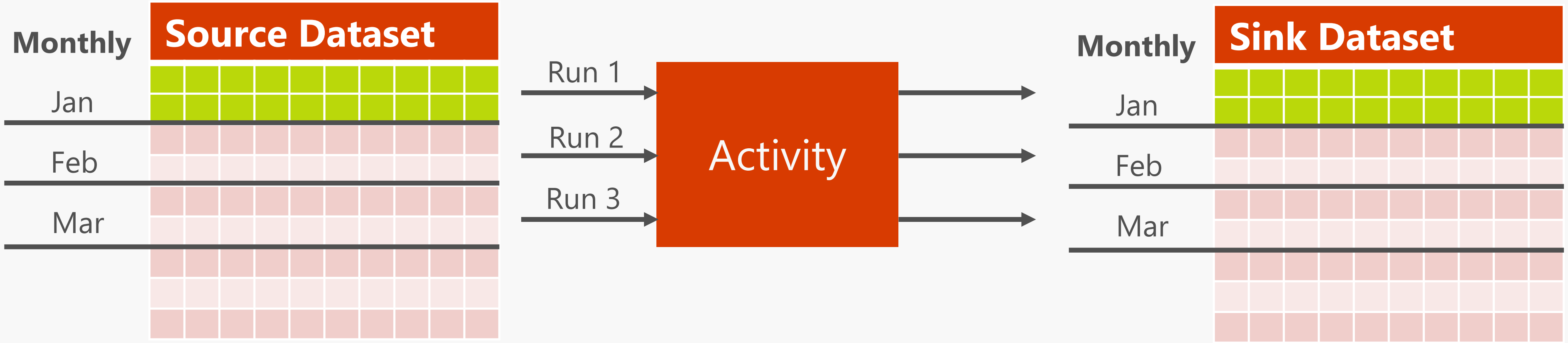
Linked Service

Data Store

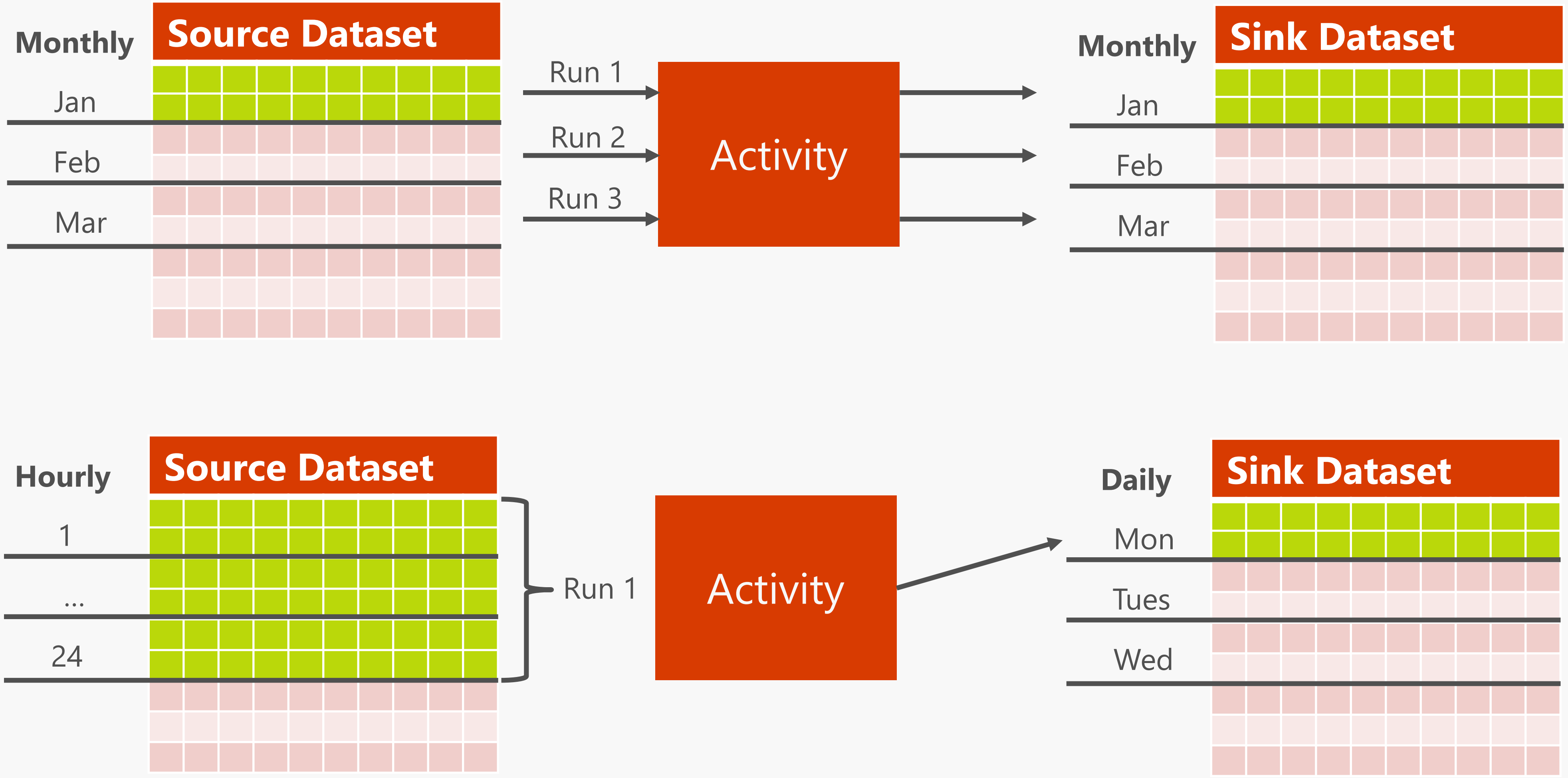
Data Factory Concepts



Scheduling and Execution



Scheduling and Execution

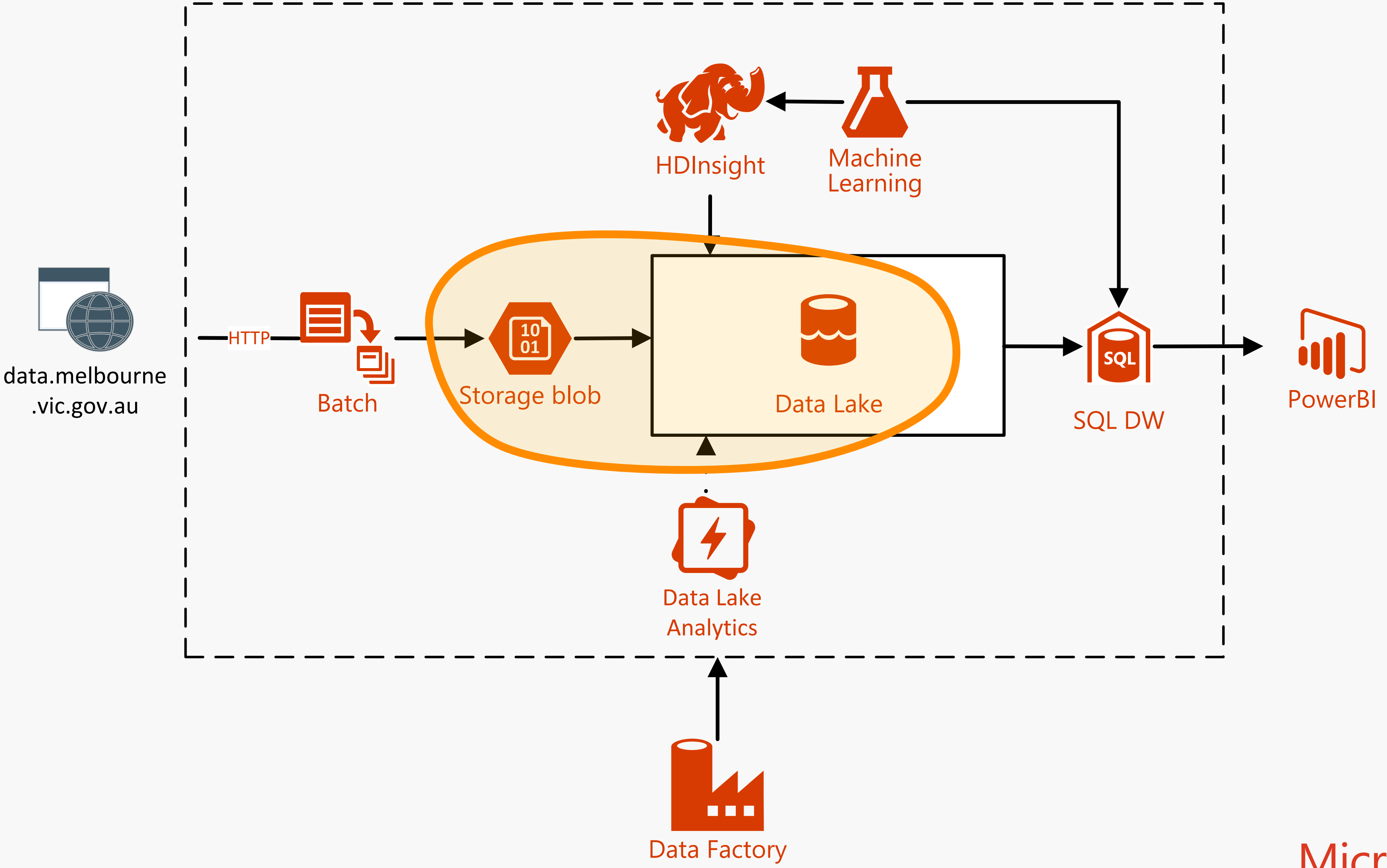


Microsoft Ignite
Australia 2017

Demo: Copy Activity

Blob to Data Lake Store

Demo: Copy Activity



Microsoft Ignite

Data Factory Design Considerations

- Ideal for time series data
- Design repeatable activity windows
- Handle ‘late’ / out-of-order runs
- Try to finalize pipeline schedules in advance
- No first class support for event driven pipelines and control flow activities
  - But with workarounds ☺

Agenda

Design

Lamda Architecture
Intro to Data Factory
Considerations

Build

Data Movement
Data Transformation

Manage

Monitor pipeline health
Developer tools

Compare

Data Factory vs Oozie

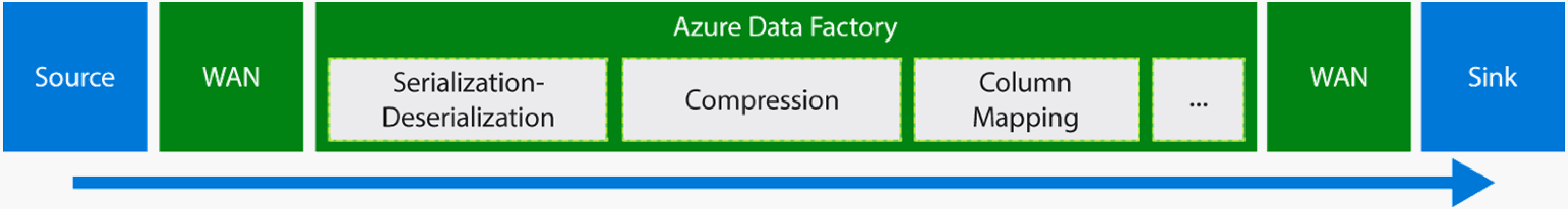
Microsoft Ignite
Australia 2017

Build: Data Movement

Microsoft

Data Movement Activity: Copy

Cloud to Cloud



Source

WAN

Azure Data Factory

Serialization-Deserialization

Compression

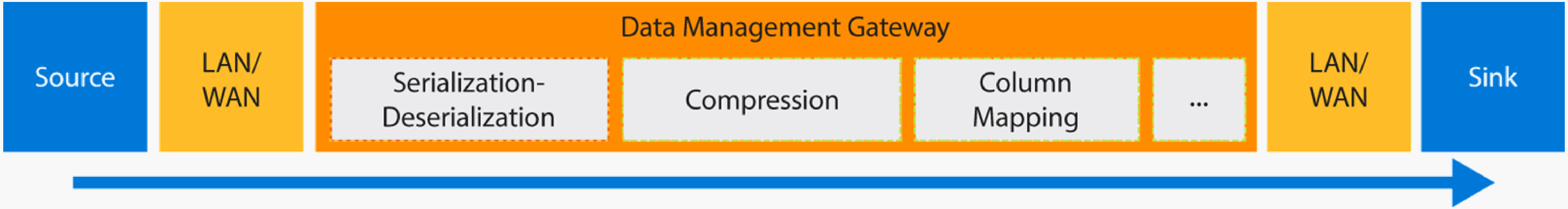
Column Mapping

…

WAN

Sink

On-premises to Cloud



Source

LAN/ WAN

Data Management Gateway

Serialization-Deserialization

Compression

Column Mapping

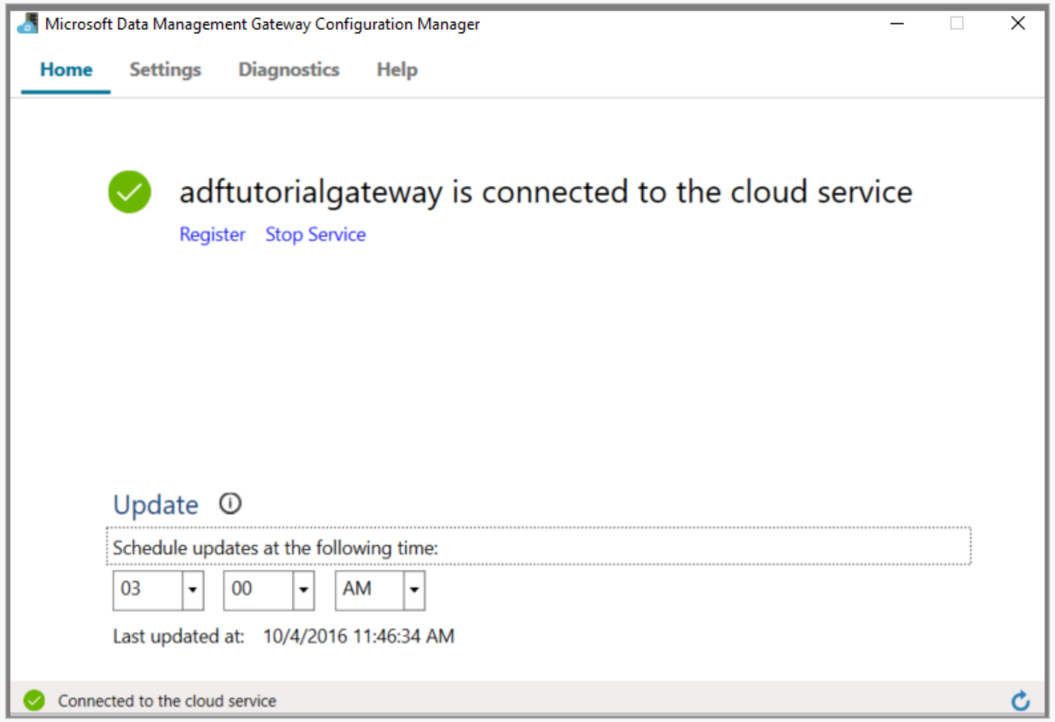
…

LAN/ WAN

Sink

Data Management Gateway

Client agent installed on-premises environment to copy data between cloud and on-premises data stores



Microsoft Data Management Gateway Configuration Manager

✓ adftutorialgateway is connected to the cloud service

Update

Schedule updates at the following time:

03 | 00 | AM

Last updated at: 10/4/2016 11:46:34 AM

✓ Connected to the cloud service

Microsoft Ignite

Supported Data Stores

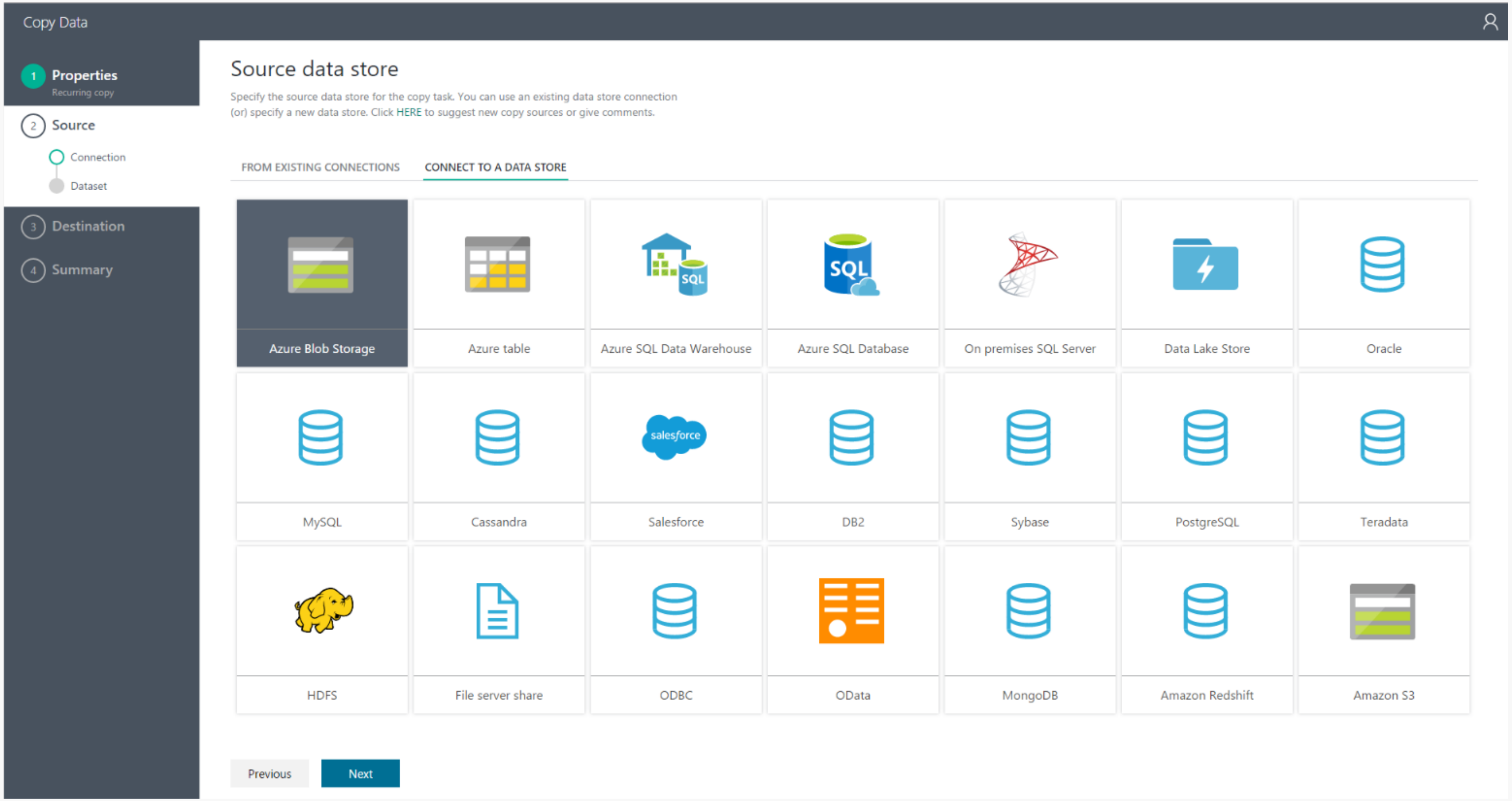
Category	Data store	Source	Sink
Azure	Azure Blob storage	✓	✓
	Azure Data Lake Store	✓	✓
	Azure SQL Database	✓	✓
	Azure SQL Data Warehouse	✓	✓
	Azure Table storage	✓	✓
	Azure DocumentDB	✓	✓
	Azure Search Index	✓	✓
Databases	SQL Server*	✓	✓
	Oracle*	✓	✓
	MySQL*	✓	
	DB2*	✓	
	Teradata*	✓	
	PostgreSQL*	✓	
	Sybase*	✓	
	Cassandra*	✓	
	MongoDB*	✓	
	Amazon Redshift	✓	

Category	Data store	Source	Sink
File	File System*	✓	✓
	HDFS*	✓	
	Amazon S3	✓	
	FTP	✓	
Others	Salesforce	✓	
	Generic ODBC*	✓	
	Generic OData	✓	
	Web Table (table from HTML)	✓	
	GE Historian*	✓	

Data stores with * can be on-premises or on Azure IaaS, and require you to install Data Management Gateway on an on-premises/Azure IaaS machine.

Microsoft Ignite

Copy Wizard



Copy Data

1 Properties
Recurring copy

2 Source
Connection
Dataset

3 Destination

4 Summary

Source data store

Specify the source data store for the copy task. You can use an existing data store connection (or) specify a new data store. Click HERE to suggest new copy sources or give comments.

Considerations

- Copy service typically runs at region closest to sink
- Performance & Tuning
  - Use Parallel Copy if you have several small files
  - Increase Cloud data movement units (DMUs)
  - Staged copy and Compression
  - Column mapping and binary copy

Microsoft Ignite
Australia 2017

Build: Data Transformation

Microsoft

Data Transformation Activities

- Stored Procedure Activity
- HDInsight Activities
- Data Analytics U-SQL Activity
- Machine Learning Activities
- .NET Custom Activity

Stored Procedure Activity

Azure SQL Database

Azure SQL Datawarehouse

Use Polybase for loading large datasets

SQL Server Database (on-premises or IAAS)

Needs Data Management Gateway

Polybase and CTAS

Polybase to access non-relational sources

‘CREATE TABLE AS’ is fully parallelized operation to create and load tables from a SELECT statements

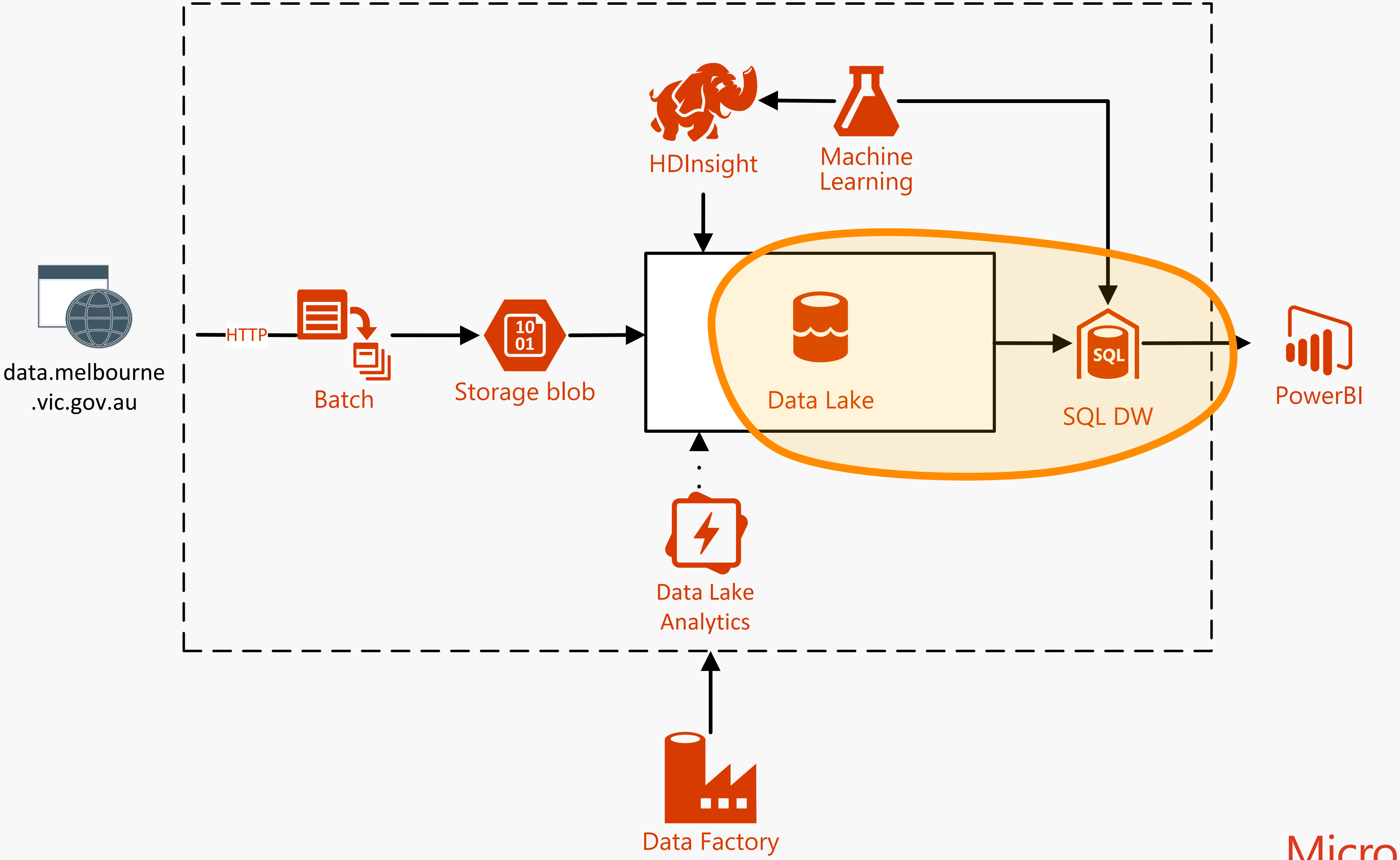
Super-charge SELECT INTO statement

Microsoft Ignite
Australia 2017

Demo: Stored Proc Activity

Loading data to SDW via Polybase

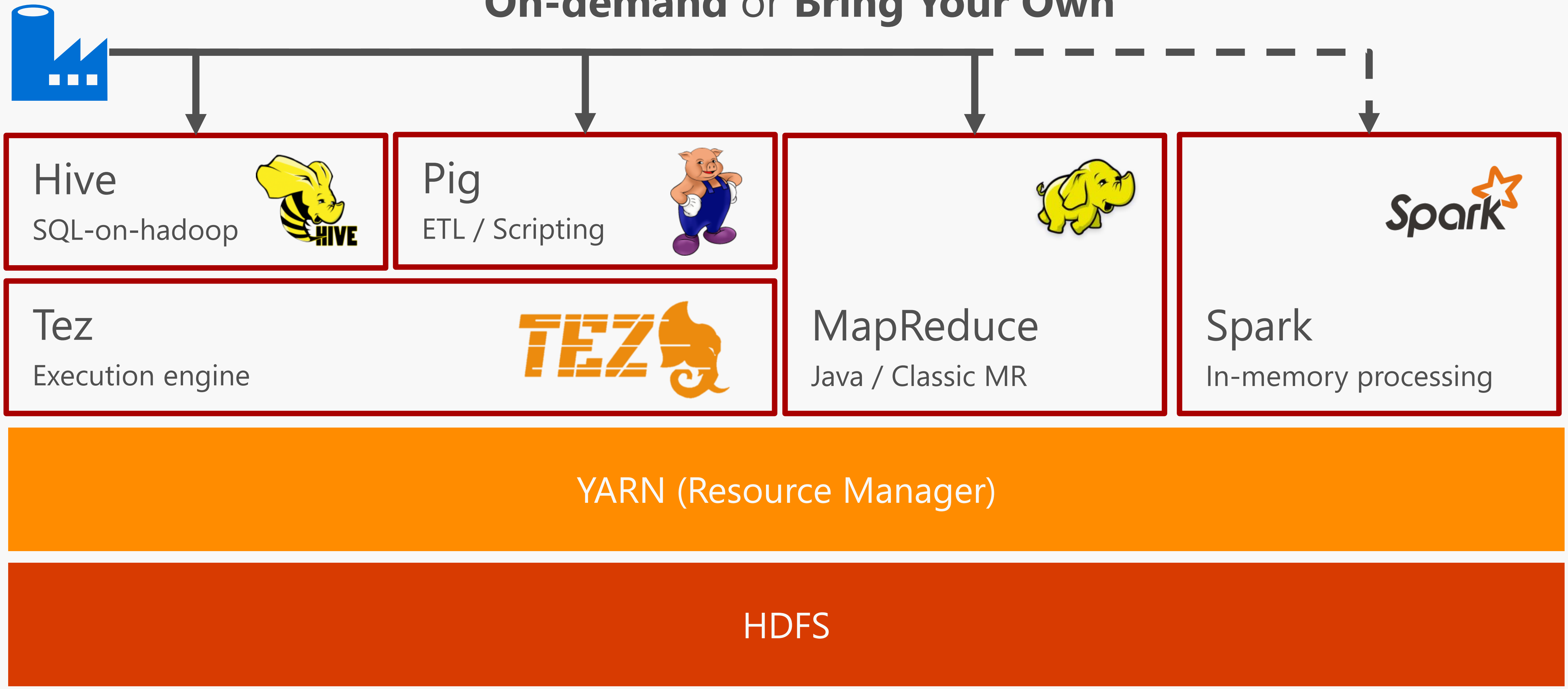
Demo: Stored Procedure Activity



Microsoft Ignite

HDInsight Activities

On-demand or **Bring Your Own**



Hive
SQL-on-hadoop

Pig
ETL / Scripting

Tez
Execution engine

MapReduce
Java / Classic MR

Spark
In-memory processing

YARN (Resource Manager)

HDFS

Hive Activity Advanced Properties

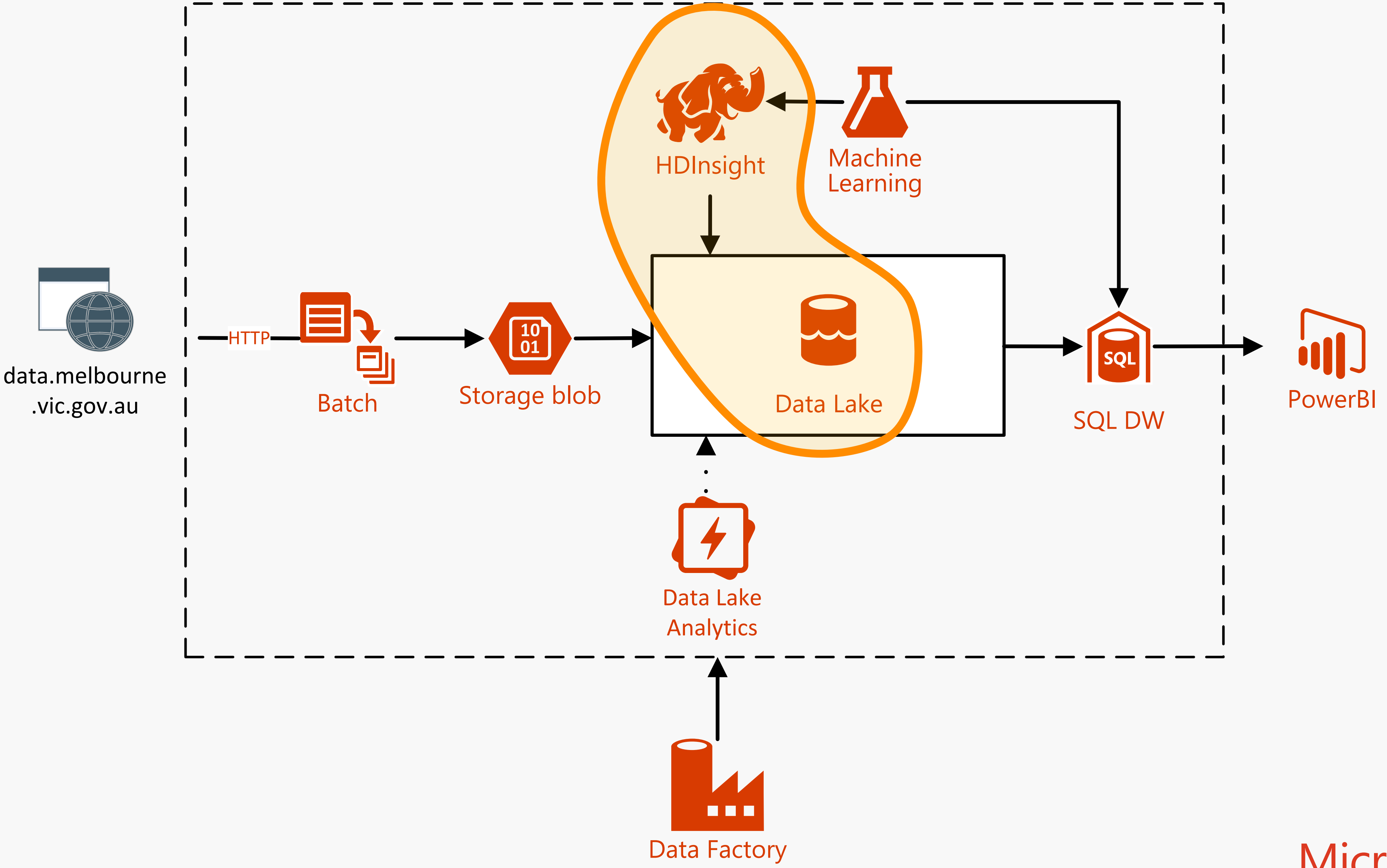
Property	Hadoop config
coreConfiguration	core-site.xml
hBaseConfiguration	hbase-site.xml
hdfsConfiguration	hdfs-site.xml
hiveConfiguration	hive-site.xml
mapReduceConfiguration	mapred-site.xml
oozieConfiguration	oozie-site.xml
stormConfiguration	storm-site.xml
yarnConfiguration	yarn-site.xml

Microsoft Ignite
Australia 2017

Demo: Hive Activity

Feature engineering

Demo: Hive Activity



Microsoft Ignite

U-SQL Activity

Data Lake Analytics is a distributed analytics service w/ federated access across Azure data stores

Comes with **U-SQL**

Big data query language that combines SQL and C#

Pay per job (no cluster needed)

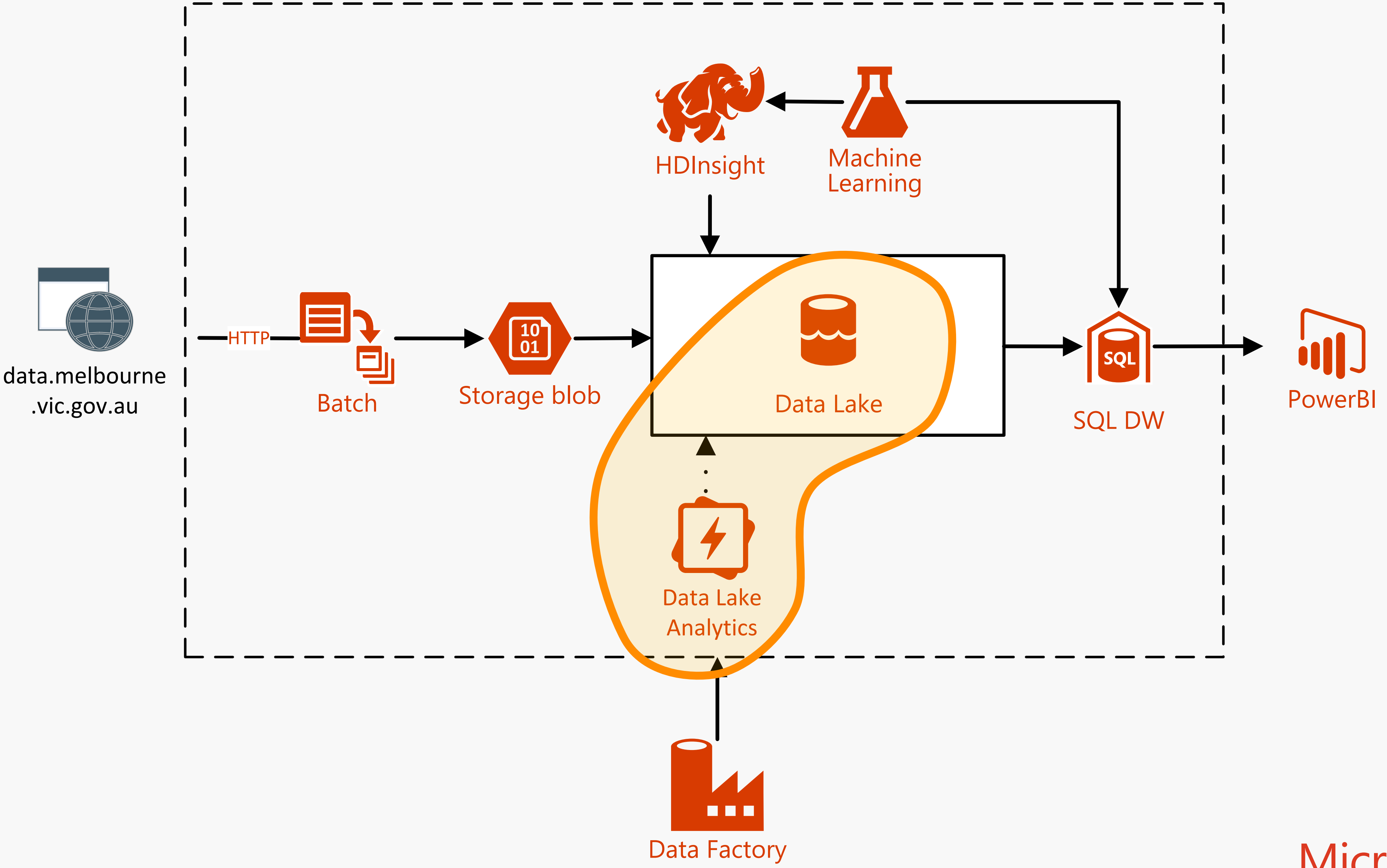
Microsoft Ignite
Australia 2017

Demo: U-SQL Activity

Feature engineering (v2)

Microsoft

Demo: U-SQL Activity



AzureML Activities

1. Batch Execution Activity

   Scoring datasets

   Retraining models

2. Update Resource Activity

   Updating ML models

Batch Execution Activity

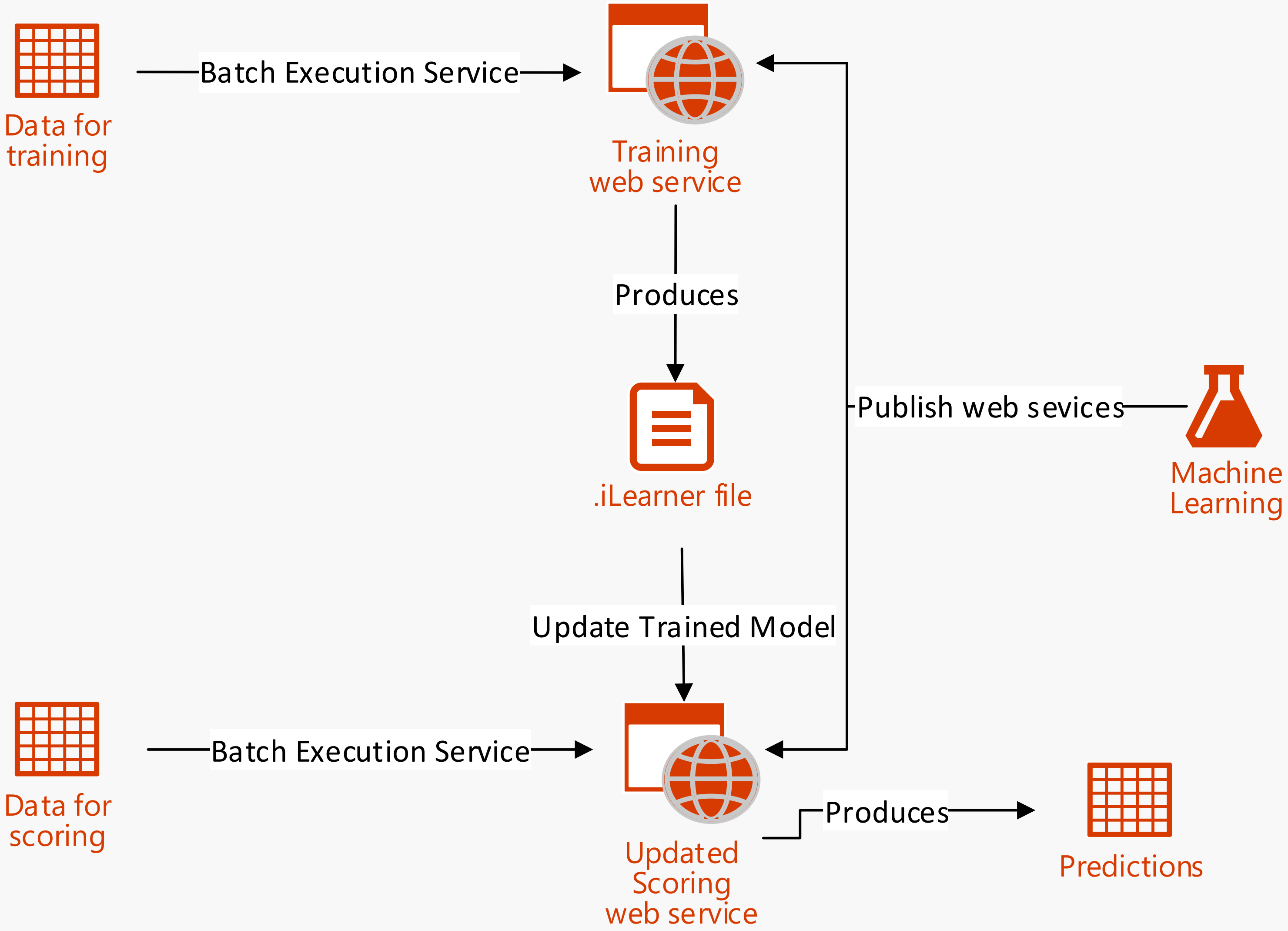
1. Use Web service inputs/outputs

   Pass datasets as inputs / outputs of web service
   Need to specify single file name as inputs

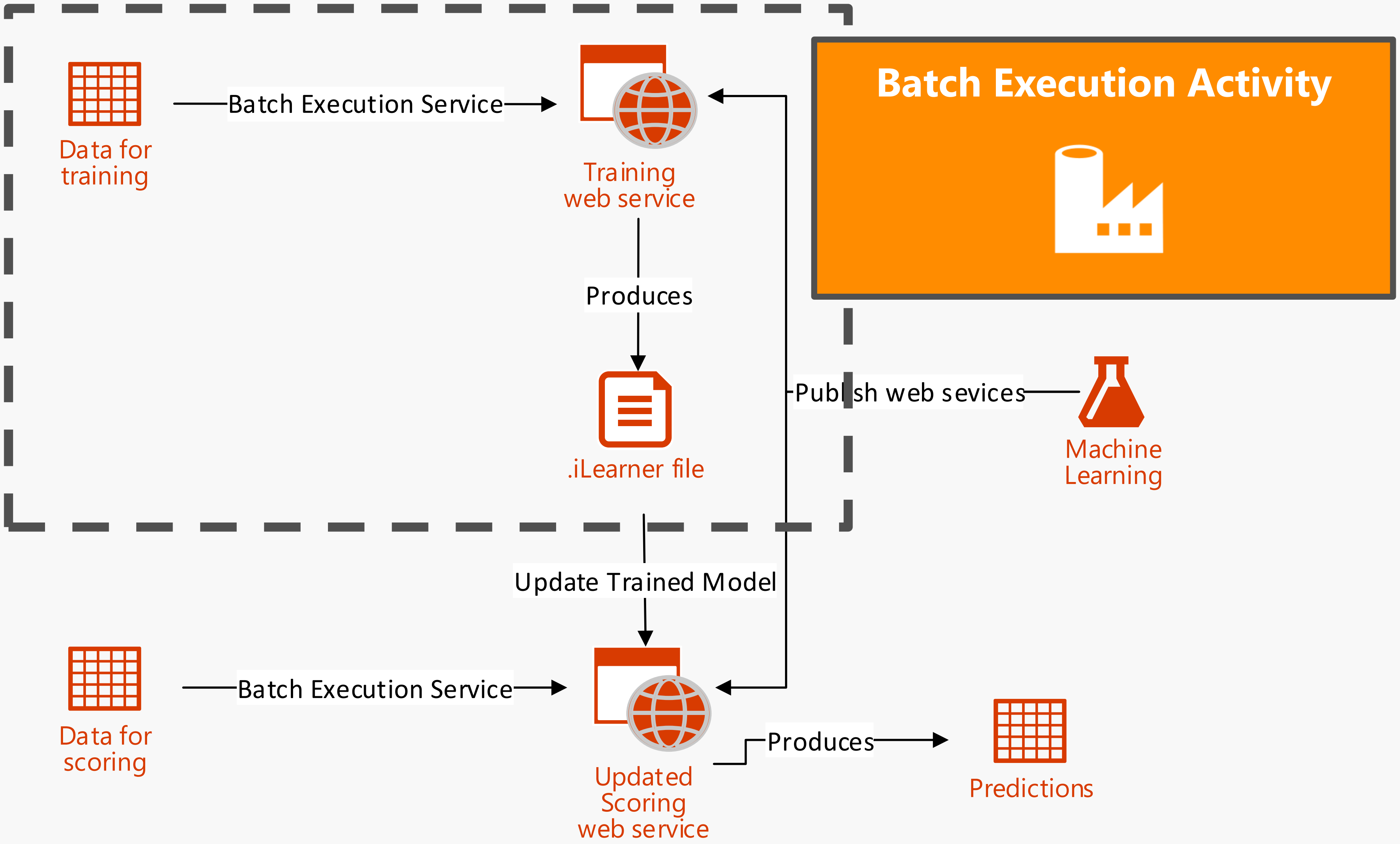
2. Use AzureML import/export modules

   Pass parameters (globalParameters) to specify reader/writer properties
   Can pass a directory as input data to readers

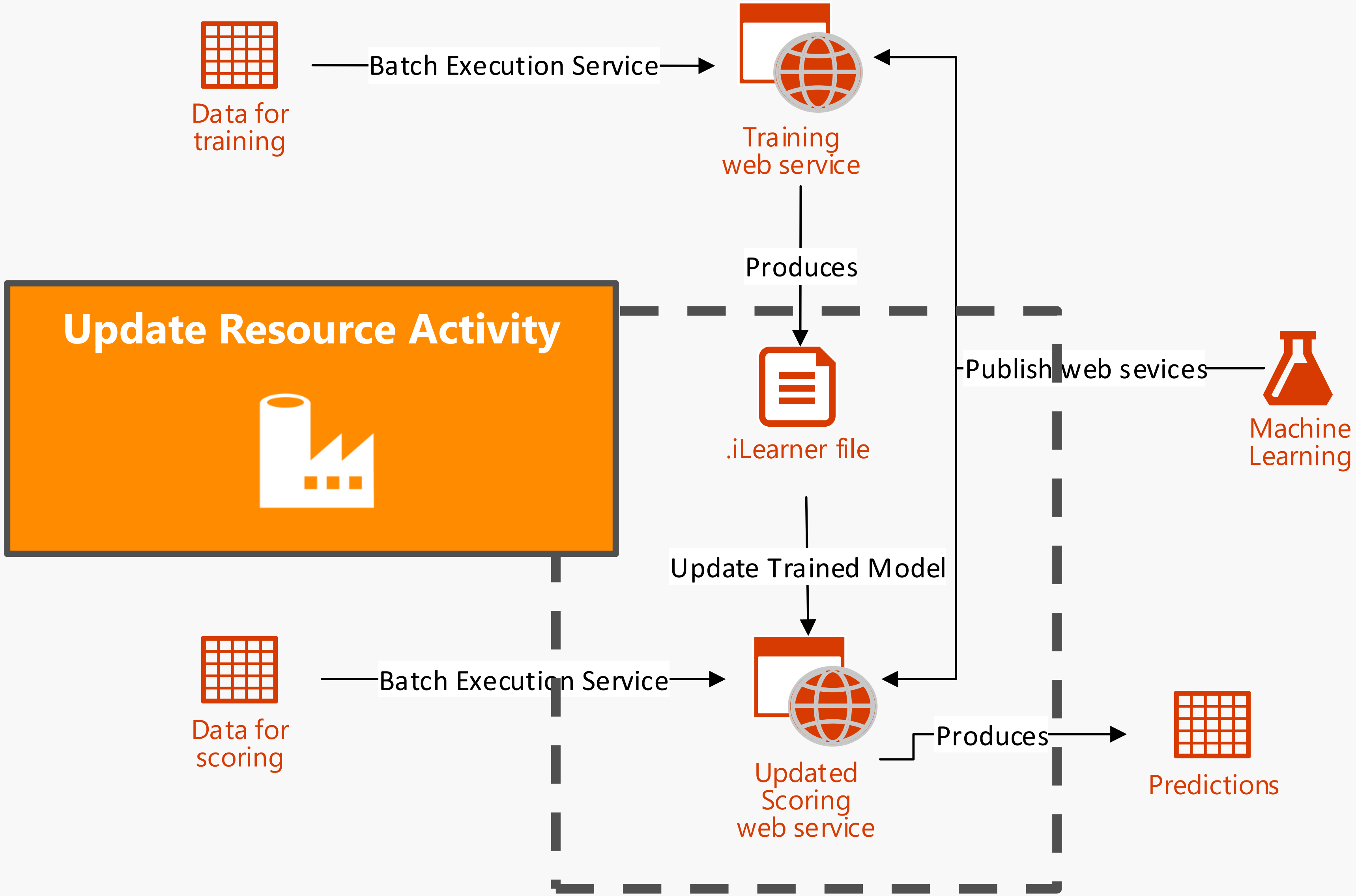
Retraining your model



Retraining your model



Retraining your model



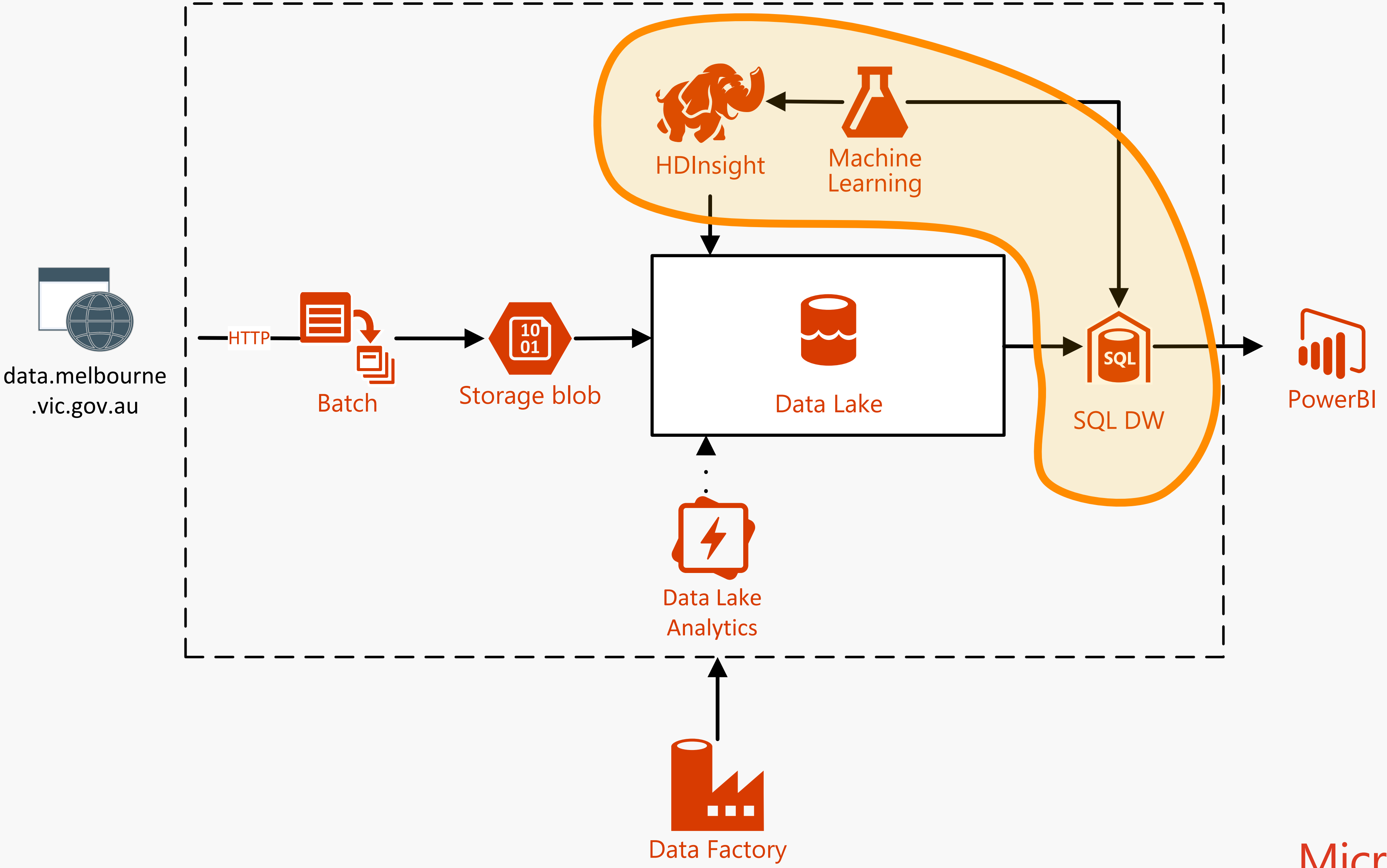
Microsoft Ignite
Australia 2017

Demo: AzureML Batch Execution Activity

Predicting Tram load

Microsoft

Demo: AzureML Batch Scoring Activity



.NET Custom Activity

Use if data source/sinks not supported by ADF

Compute: **Azure Batch** or **HDInsight**

.NET Custom Activity

```
public class ApiDownloadActivity : IDotNetActivity
{
      public IDictionary<string, string> Execute(
              IEnumerable<LinkedService> linkedServices,
              IEnumerable<Dataset> datasets,
              Activity activity,
              IActivityLogger logger)
      {
           //your code

      }
}
```

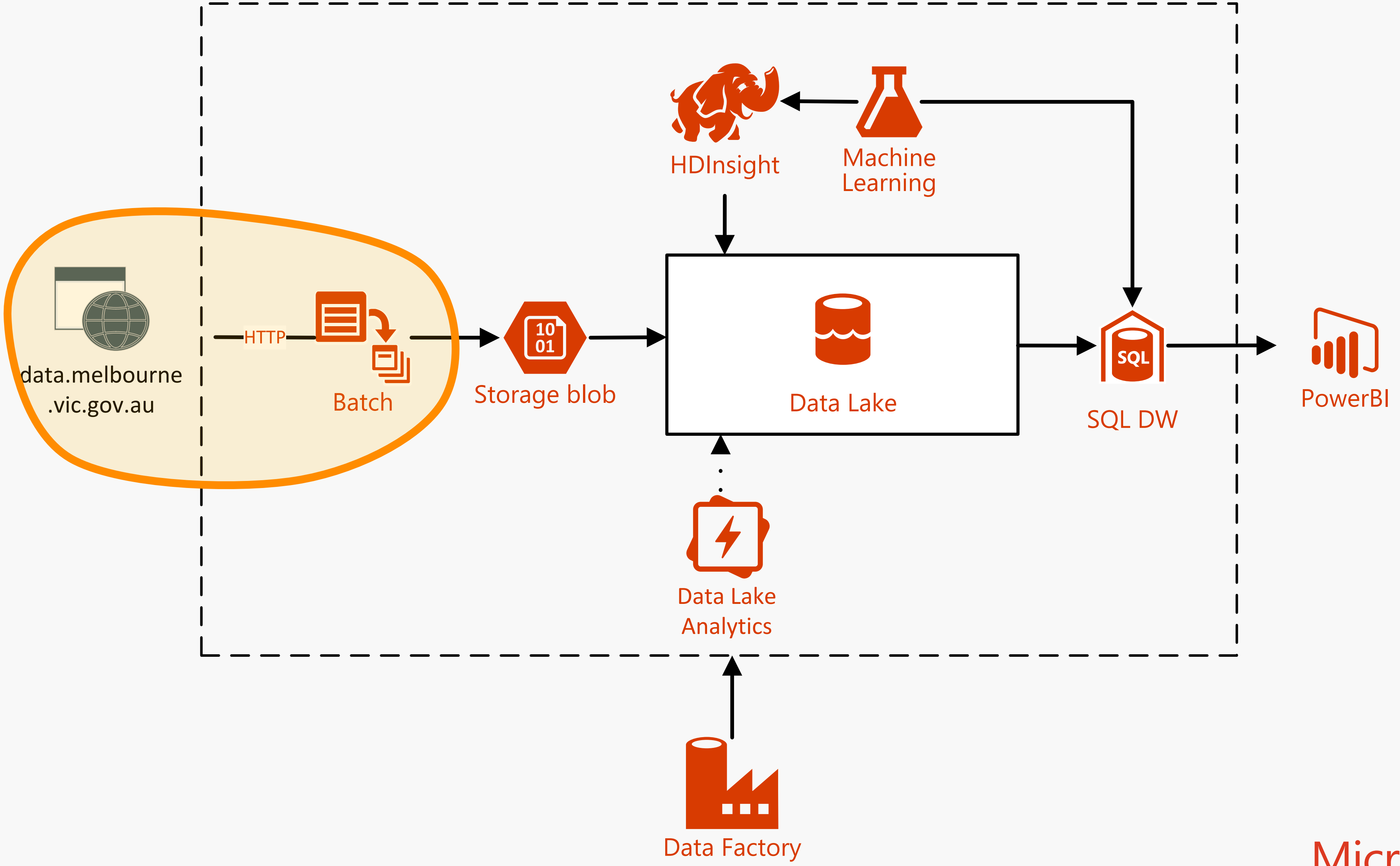
Microsoft Ignite
Australia 2017

Demo: .NET Custom Activity

Calling an API with Data Factory

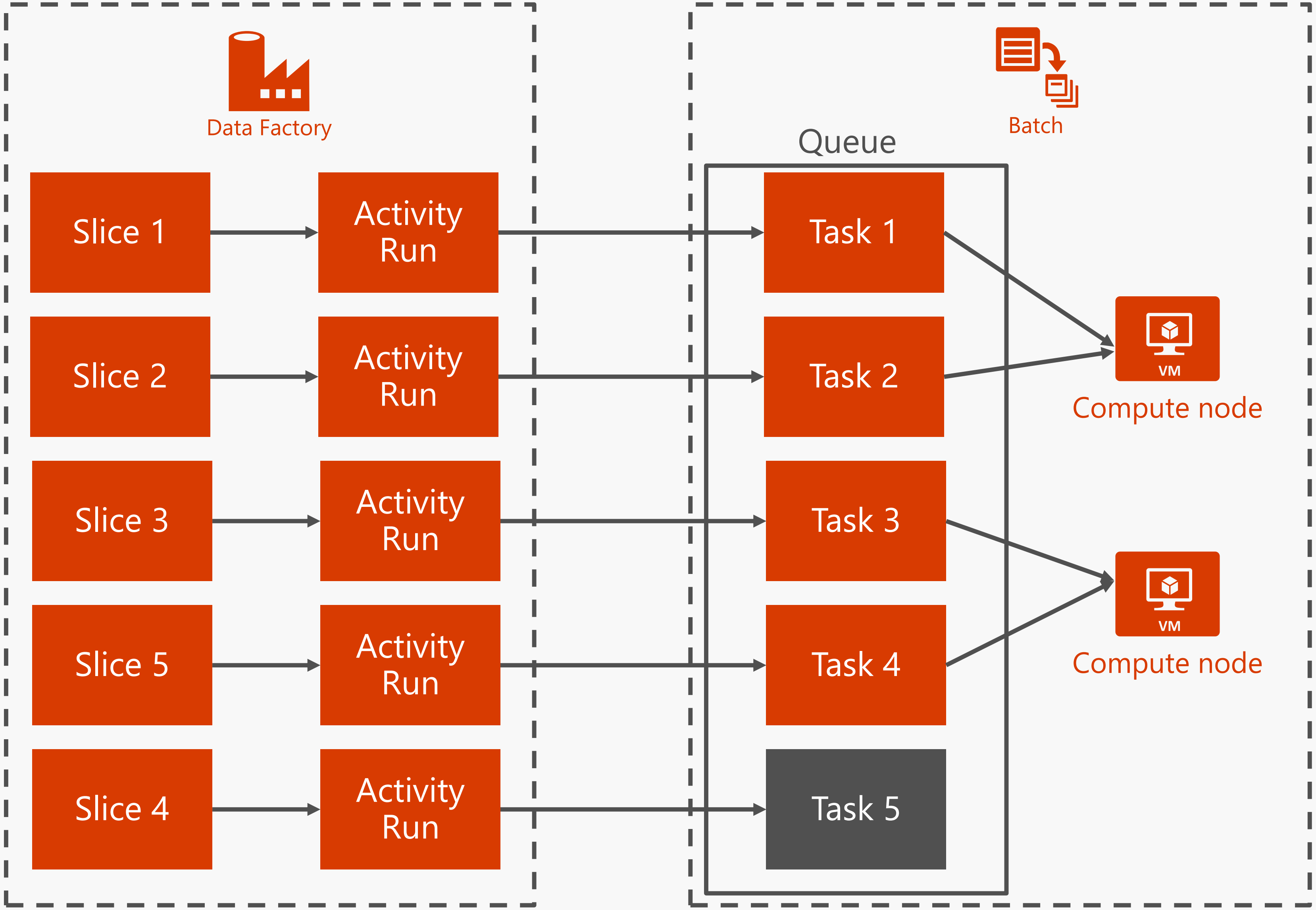
Microsoft

Demo: Custom Activity on Batch



Microsoft Ignite

.NET Custom Activity on Batch



Microsoft Ignite

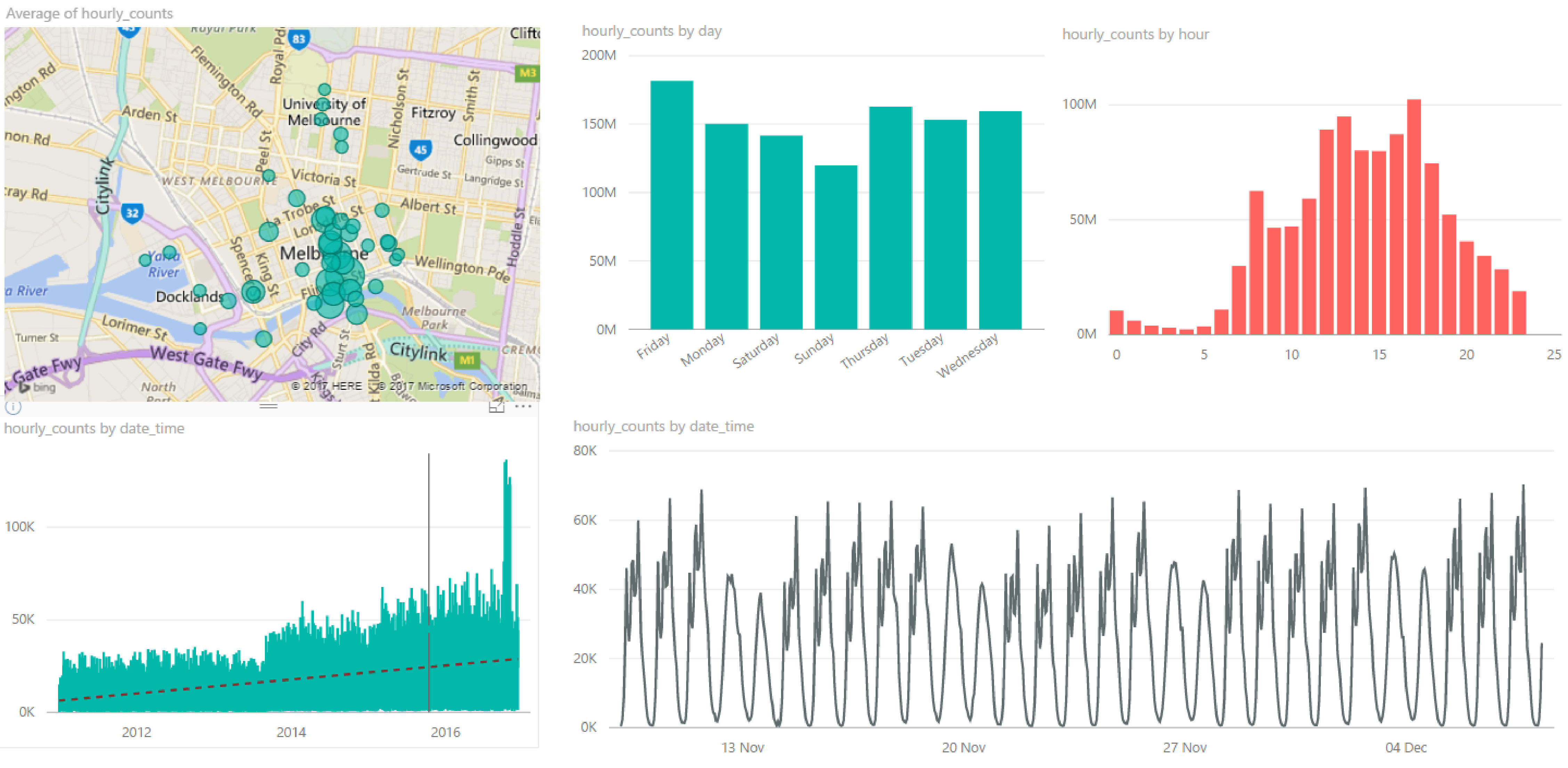
Microsoft Ignite
Australia 2017

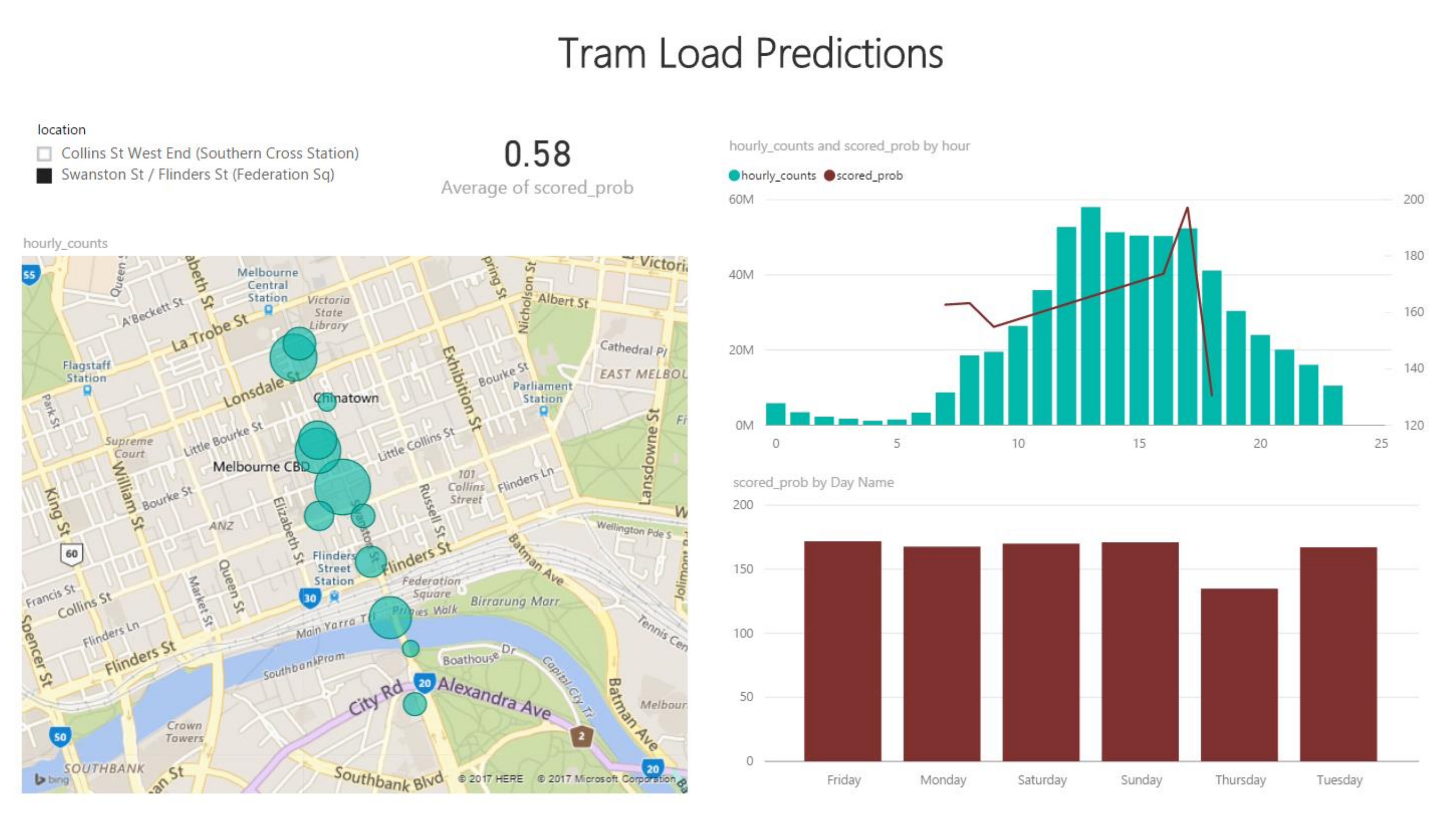
PowerBI Dashboard

Predicting tram load from pedestrian traffic

Microsoft

Melbourne Pedestrian Foot Traffic





Tram Load Predictions

location

□ Collins St West End (Southern Cross Station)

■ Swanston St / Flinders St (Federation Sq)

0.58

Average of scored_prob

Agenda

Design

Big data pipelines
Lamda Architecture
Data Factory Concepts

Build

Data Movement
Data Transformation

Manage

Monitor pipeline health
Developer tools

Compare

Data Factory vs Oozie

Microsoft Ignite
Australia 2017

Monitor and Manage

Microsoft

Monitor and Manage Dashboard

Pause and resuming pipelines

Creating alerts

Re-running failed pipelines

Microsoft Ignite
Australia 2017

Demo: Monitor and Manage

Microsoft

Developer tools

Azure Portal

VS Data Factory Extension

PowerShell

Azure Resource Manager Templates

C# SDK

Agenda

Design

- Big data pipelines
- Lamda Architecture
- Data Factory Concepts

Build

- Data Movement
- Data Transformation

Manage

- Monitor pipeline health
- Developer tools

Compare

- Data Factory vs Oozie

Microsoft Ignite
Australia 2017

Compare: Data Factory vs Oozie

Microsoft

What is Oozie?

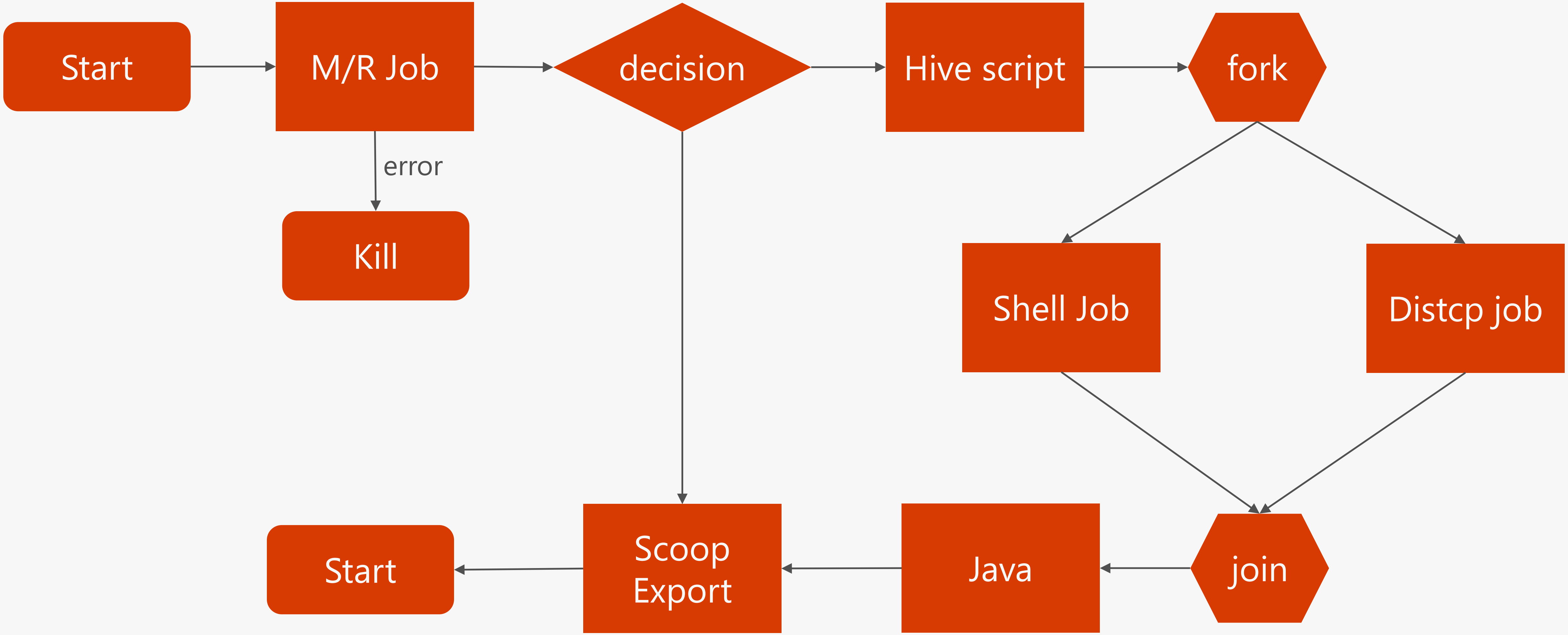
Oozie is a workflow scheduler system to manage Apache Hadoop jobs

De-facto workflow scheduler of the Hadoop stack

Out-of-the-box integration with Hadoop jobs

- Java map-reduce, Streaming map-reduce, Pig, Hive, Sqoop and shell

Sample Oozie workflow



Start

M/R Job

error

Kill

decision

Hive script

fork

Shell Job

Distcp job

join

Java

Scoop Export

Start

Data Factory vs Oozie

	Data Factory	Oozie
Sources and Sinks	Azure Hadoop Stack Numerous 3rd party	Hadoop Stack Relational through Sqoop Shell, Distcp
Hybrid pipelines	Data Management Gateway	None
Tooling	Visual Studio, Portal, PS	Hue, Eclipse plugin, 3rd party
Extensibility	C# Custom Activity	Java Action Node
Control flow	Limited	Fork and Join, Decision Control, Kill (on error)
Event-based	No first class support. Some workarounds	Dataset polling
Performance	Designed for big data workloads	Designed for big data workloads
Maturity	Preview in Oct 2014, GA in August 2015	Developed since 2008, Open sourced since 2010

Overall Summary

- Data Factory

…can orchestrate data pipelines at scale

…has tight integration with Azure’s big data PAAS offerings and with variety of 3rd party source systems

…offers data Movement and Data Transformation activities

…first choice orchestrator for **Azure** services

Session evaluation

Complete your session evaluation on MyIgnite
for your chance to **WIN** one of many daily prizes.

(image of prizes tbc)

Continue your Ignite learning path

Visit Channel 9 to access a wide range of Microsoft training and event recordings **https://channel9.msdn.com/**

Head to the TechNet Eval Centre to download trials of the latest Microsoft products **http://Microsoft.com/en-us/evalcenter/**

Visit Microsoft Virtual Academy for free online training visit **https://www.microsoftvirtualacademy.com**

Thank you

Chat with me in the Speaker Lounge
Find me @ linkedin.com/in/lacelofranco/