DevDaysAsia 2016 @ Taipei

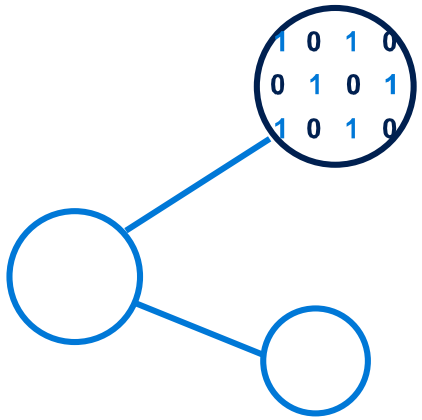
Big Data Ingestion and Storage

Darwin Schweitzer

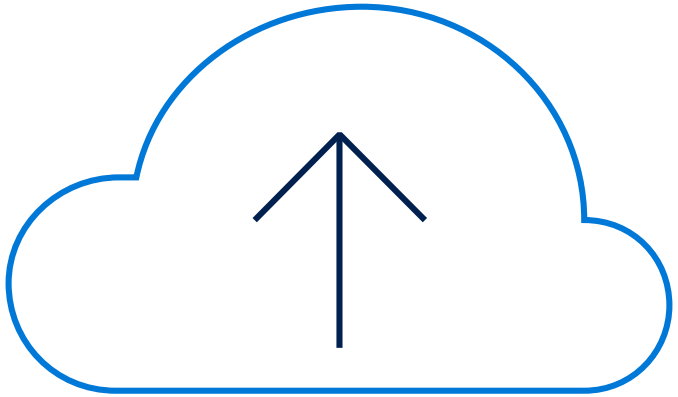
Senior Program Manager

Microsoft

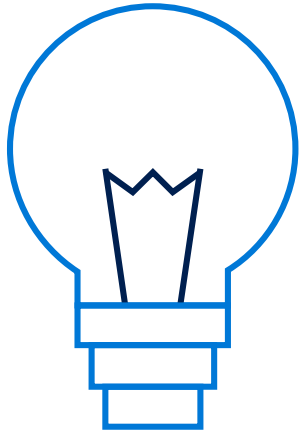
Business is being transformed by three trends



Big Data



Cloud

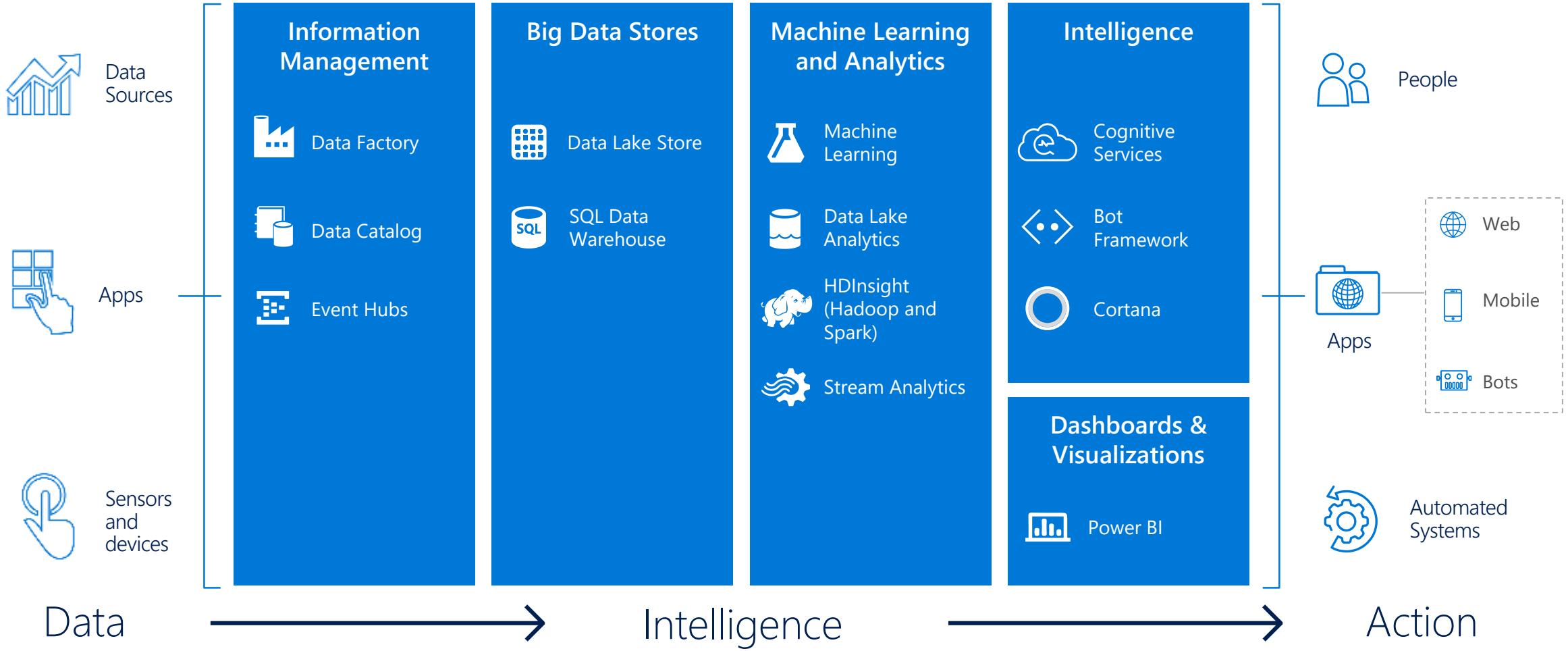


Intelligence

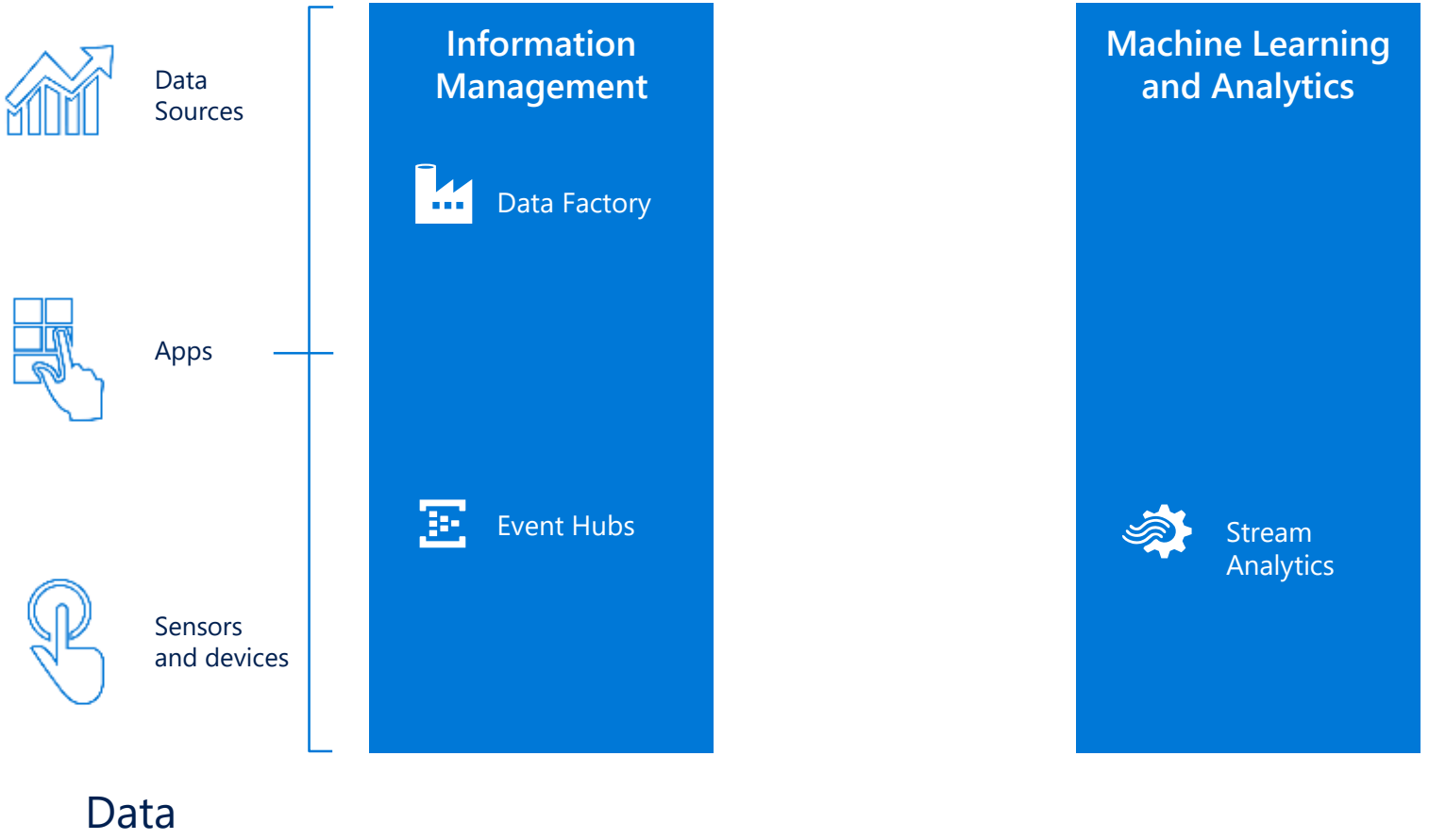
Stay ahead of the curve with Cortana Intelligence Suite

Data → Intelligence → Action

Easily turn data into intelligent action



Big Data Ingestion

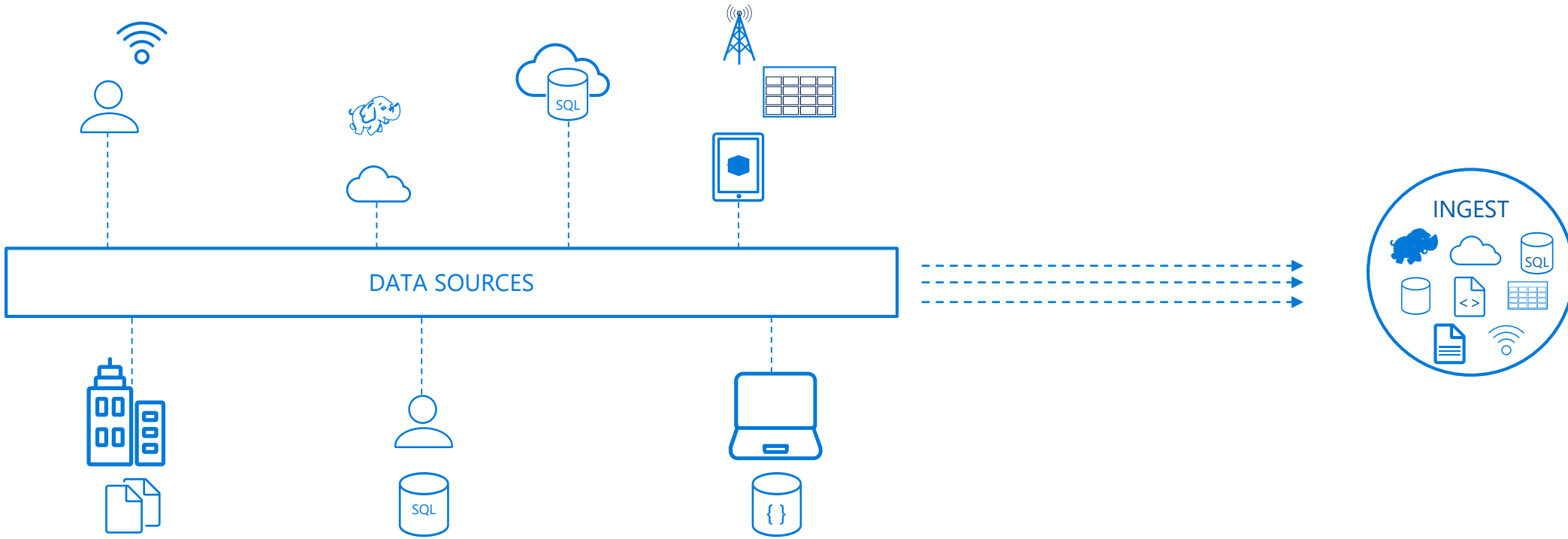


Compose and orchestrate data services at scale

Information Management

Data Factory

Event Hubs



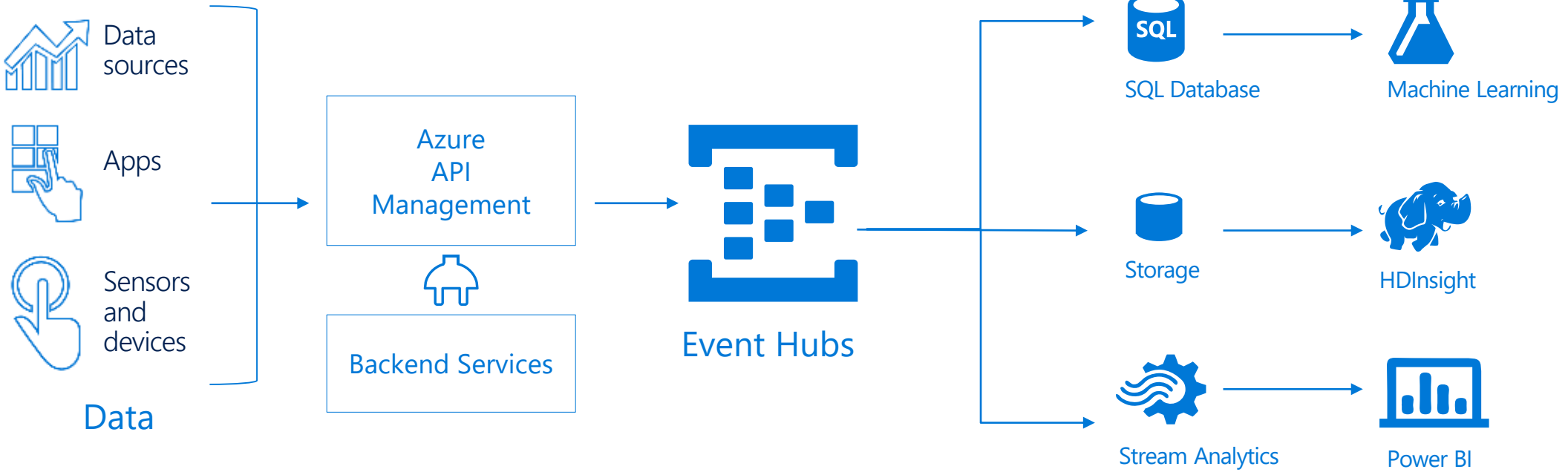
- Create, schedule, orchestrate, and manage data pipelines
- Visualize data lineage
- Connect to on-premises and cloud data sources
- Monitor data pipeline health
- Automate cloud resource management
- Move relational data for Hadoop processing
- Transform with Hive, pig, or custom code

Ingest events from websites, apps and devices at cloud scale

Information Management

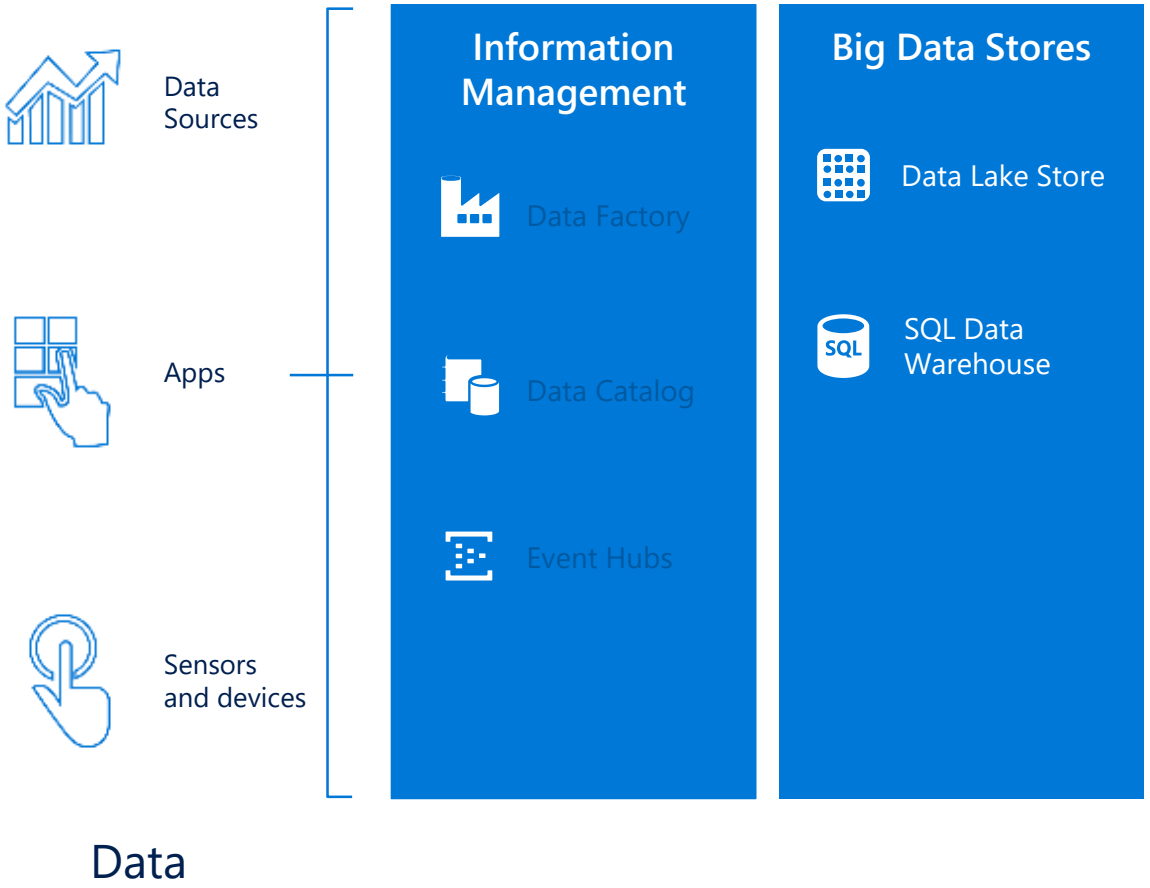
Data Factory

Event Hubs



- Log millions of events per second in near real time
- Connect devices using flexible authorization and throttling
- Use time-based event buffering
- Get a managed service with elastic scale
- Get a managed service with elastic scale
- Reach a broad set of platforms using native client libraries
- Pluggable adapters for other cloud services

Big Data Stores

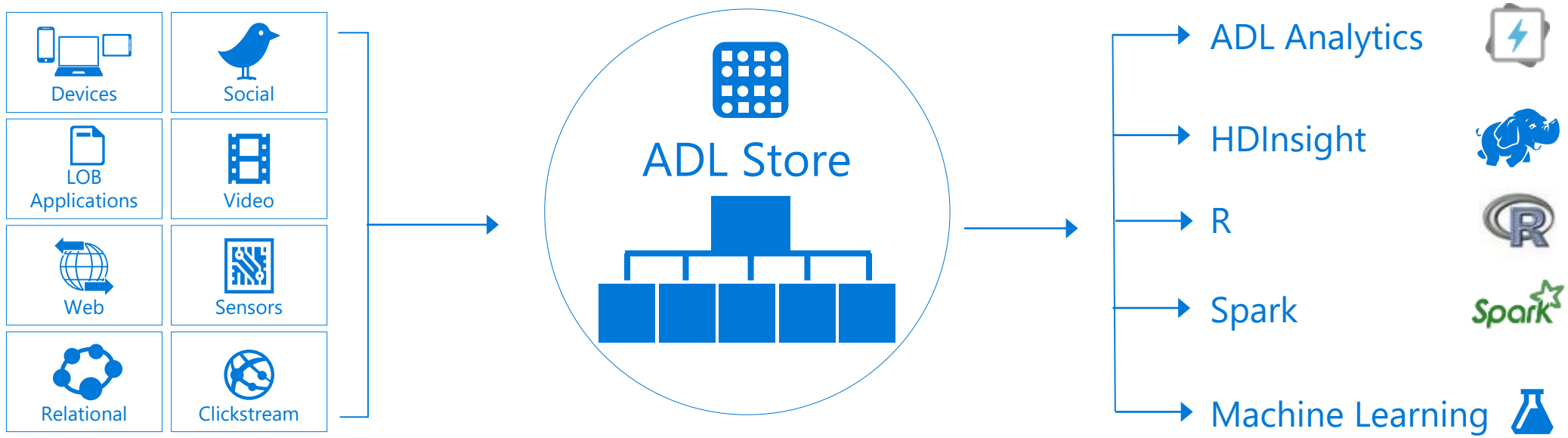


Data

A hyper-scale repository for big data analytics workloads

Big Data Stores

- Data Lake Store
- SQL Data Warehouse

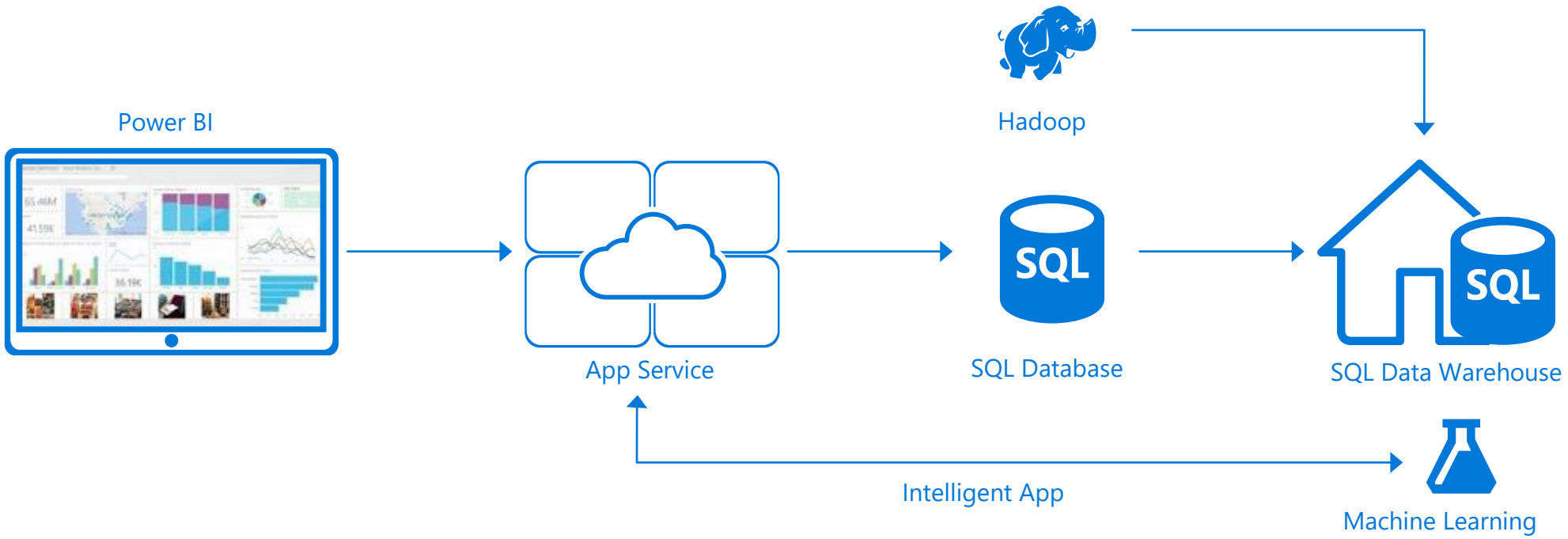


- A Hadoop Distributed File System for the cloud
- No fixed limits on file size
- No fixed limits on account size
- Unstructured and structured data in their native format
- Massive throughput to increase analytic performance
- High durability, availability, and reliability
- Azure Active Directory access control

Elastic data warehouse as a service with enterprise-class features

Big Data Stores

- Data Lake Store
- SQL Data Warehouse

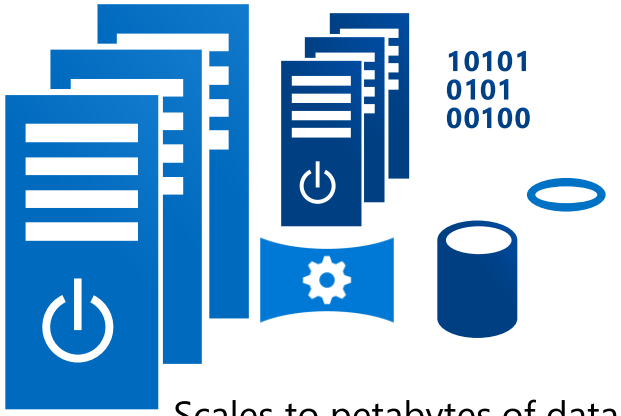


- Petabyte scale with massively parallel processing
- Independent scaling of compute and storage—in seconds
- Transact-SQL queries across relational and non-relational data
- Full enterprise-class SQL Server experience
- Works seamlessly with Power BI, Machine Learning, HDInsight, and Data Factory

Introducing Azure SQL DW Service

A relational **data warehouse-as-a-service**, fully managed by Microsoft.
Industries first **elastic** cloud data warehouse with proven SQL Server capabilities.
Support your **smallest to your largest** data storage needs.

Elastic scale & performance



Scales to petabytes of data

Massively Parallel Processing

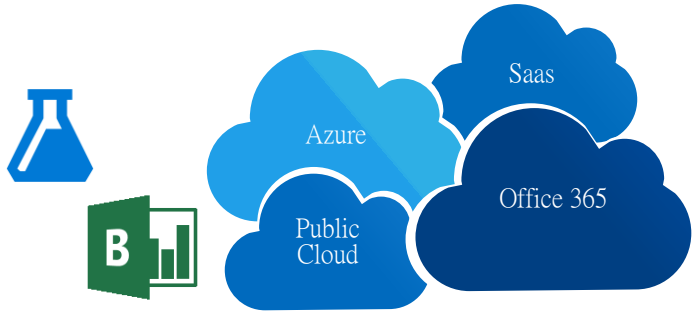
Instant-on compute scales in seconds

Query Relational / Non-Relational

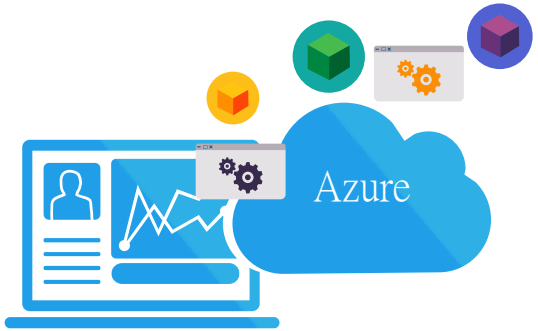
Powered by the Cloud

Get started in minutes

Integrated with Azure ML, PowerBI & ADF



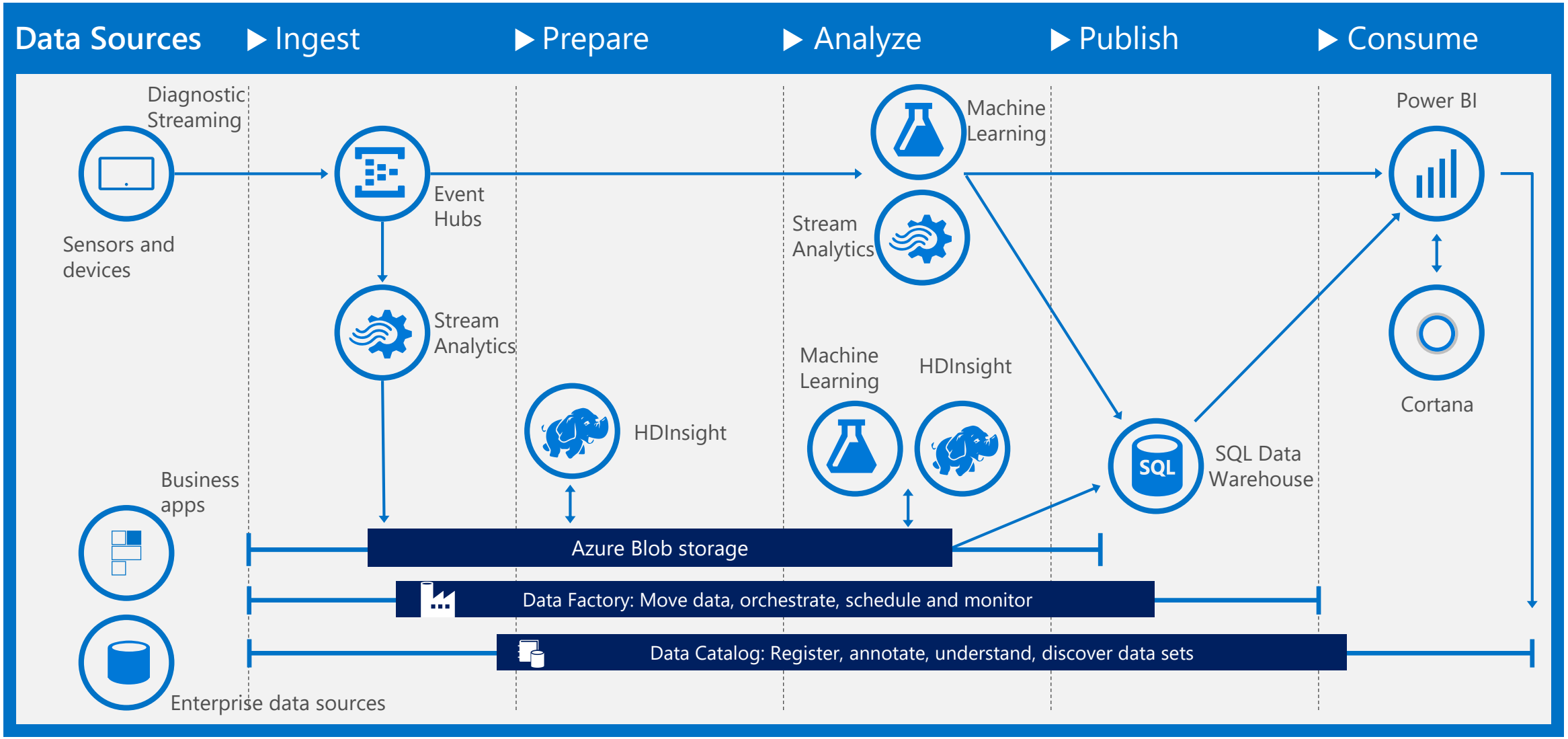
Market Leading Price & Performance



Simple billing compute & storage

Pay for what you need, when you need it with dynamic pause

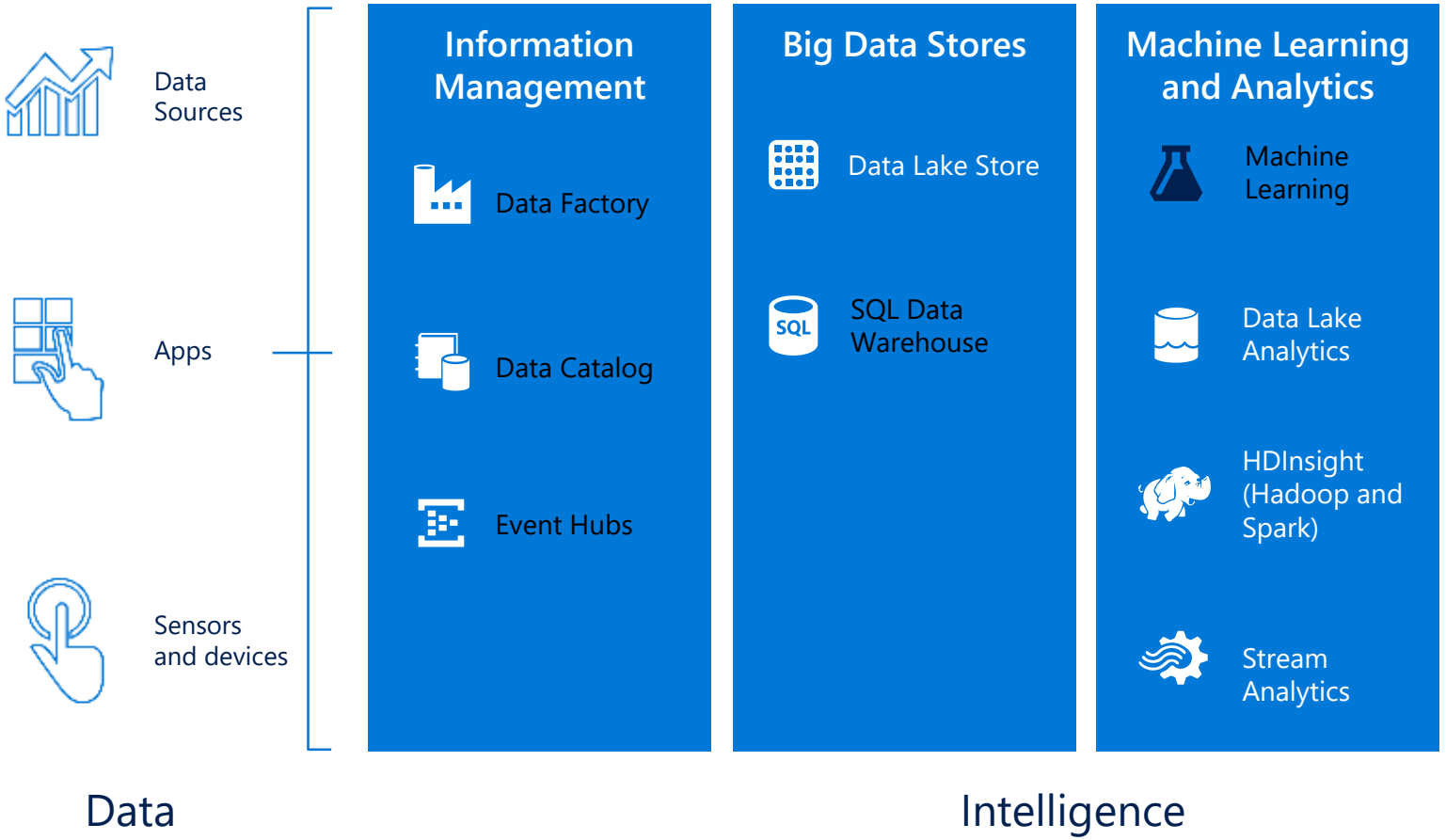
Example of Cortana Intelligence Suite in action



Demo Azure SQL Data Warehouse

Microsoft

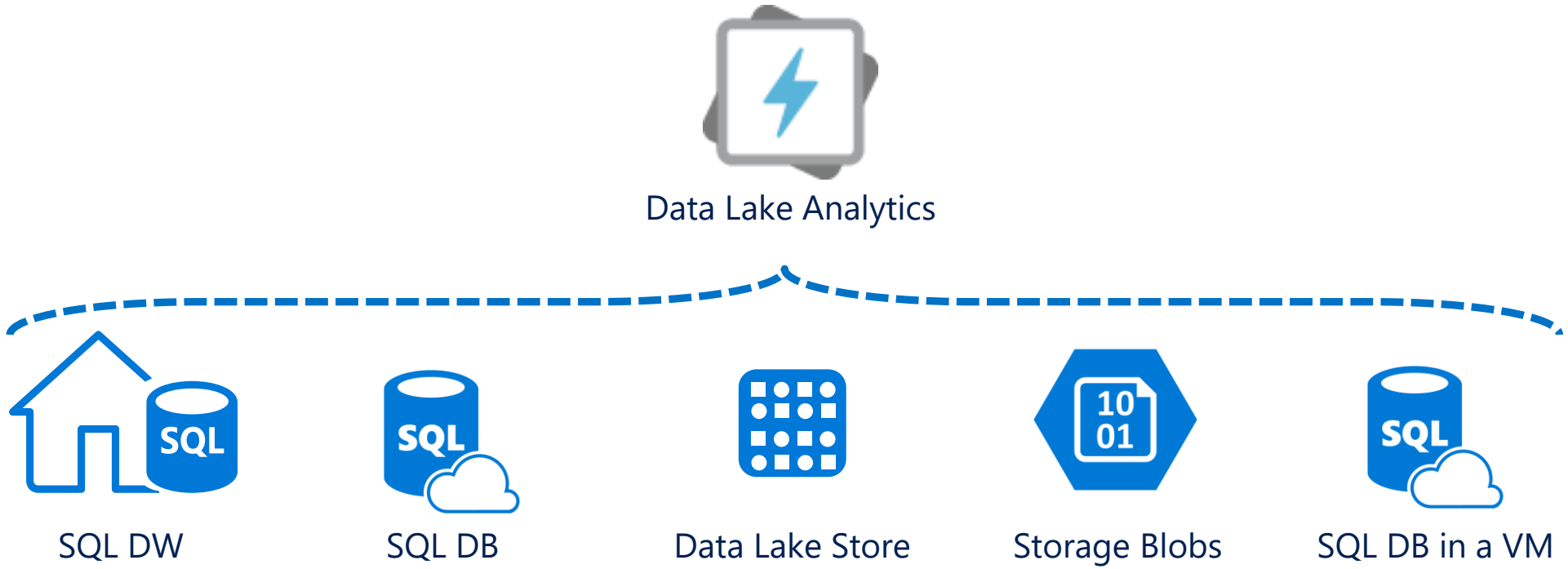
Machine Learning and Analytics



Big data analytics made easy

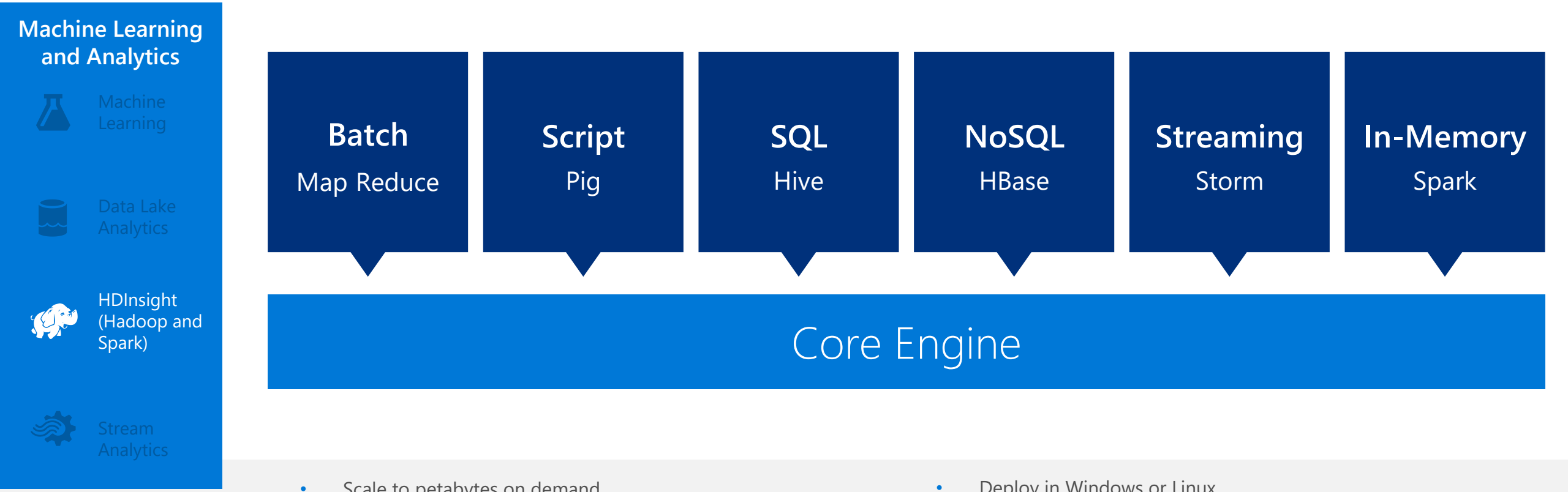
Machine Learning and Analytics

- Machine Learning
- Data Lake Analytics
- HDInsight (Hadoop and Spark)
- Stream Analytics



- Analyze data of any kind and size
- Develop faster, debug and optimize smarter
- Interactively explore patterns in your data
- No learning curve—use U-SQL, Spark, Hive, HBase and Storm
- Managed and supported with an enterprise-grade SLA
- Dynamically scales to match your business priorities
- Enterprise-grade security with Azure Active Directory
- Built on YARN, designed for the cloud

Comprehensive set of managed Apache big data projects



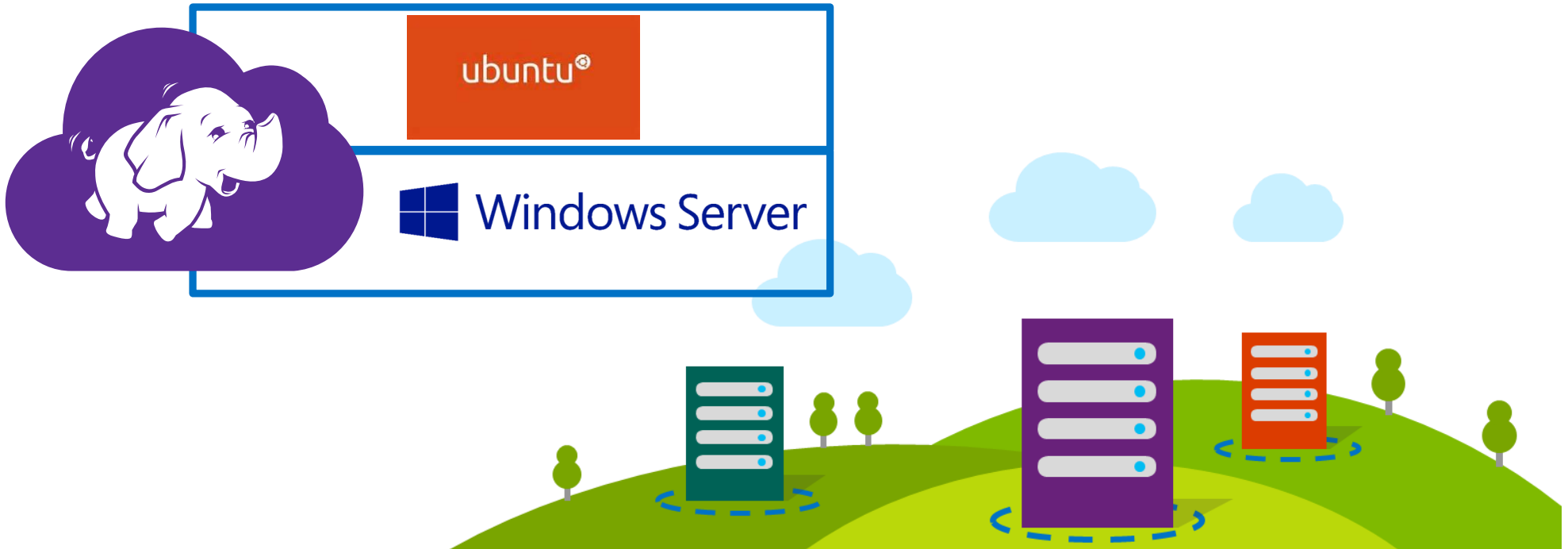
- Scale to petabytes on demand
- Process unstructured and semi-structured data
- Develop in Java, .NET, and more
- Skip buying and maintaining hardware

- Deploy in Windows or Linux
- Spin up an Apache Hadoop cluster in minutes
- Visualize your Hadoop data in Excel
- Easily integrate on-premises Hadoop clusters

Azure HDInsight running Linux

Choice of Windows or Linux clusters

Managed & supported by Microsoft
Re-use common tools, documentation, samples from Hadoop/Linux ecosystem
Add Hadoop projects that were authored on Linux to HDInsight
Easier transition from on-premises to cloud



Strata Announcements

https://blogs.technet.microsoft.com/machinelearning/2016/03/29/microsoft-makes-big-data-analytics-easier-in-the-cloud/

Strata+Hadoop WORLD

MAKE DATA WORK
MARCH 28–29, 2016: TRAINING
MARCH 29–31, 2016: CONFERENCE
SAN JOSE, CA

SCHEDULE | SPEAKERS | EXPO | EVENTS | VENUE | ABOUT | RESOURCES

Strata + Hadoop World 2016 Speakers

New speakers are added regularly. Please check back to see the latest updates agenda.

All Speakers | Sessions | Tutorials | 2-day Training

Search Speakers Joseph S

Joseph Sirosh (Microsoft), @josephsirosh
Connected brains Keynote

Joseph Sirosh is the corporate vice president of the Data Group, le
machine-learning products, as well as a talented team of engineer
developing tools and services to transform data at scale into actio

Machine Learning Blog

Microsoft Makes Big Data and Analytics Easier in the Cloud

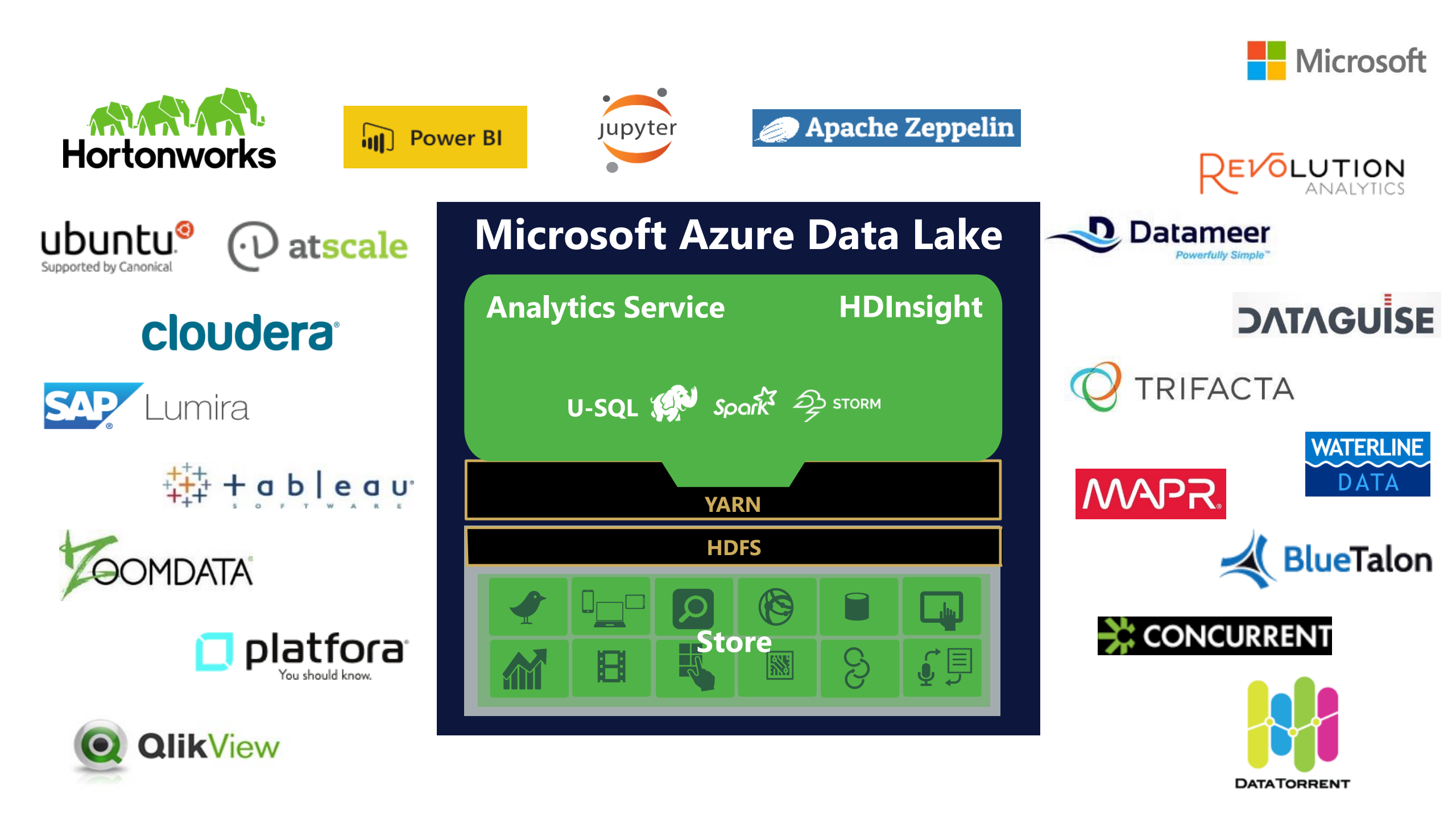
March 29, 2016 by ML Blog Team // 0 Comments

f 0 | 0 | in 0

This post is by Joseph Sirosh, Corporate Vice President of the Data Group at Microsoft.

This week I'm joining thousands of people attending Strata + Hadoop World in San Jose to explore the technology and business of big data and data science. As part of our participation in the conference, we are announcing several important investments to continue delivering on our commitment to make big data processing and analytics simpler and more accessible:

- **Advanced analytics at scale with R Server for HDInsight and the latest version of Spark for HDInsight are now available in preview**: Customers can leverage their existing R skills and reuse current code to run at scale. R Server for HDInsight offers popular scalable R algorithms and the ability to parallelize any existing R function. We are also releasing the latest version of Spark for HDInsight, which can deliver **7x** performance over MapReduce for most analytics. These capabilities give our customers the ability to train and run advanced analytics and ML models on larger datasets, and much faster than previously possible in the cloud.
- **Out-of-the-box application integration, providing easier access to popular big data apps:** Customers can now discover and deploy popular big data applications with HDInsight without any code or scripting required. Leading solutions such as Datameer Cloud offer code-free data preparation, AtScale has cloud-based OLAP BI on Hadoop, and an ecosystem of other big data applications can now be deployed alongside HDInsight.
- **Azure Data Catalog**, previously announced as a public preview will be generally available tomorrow: Data Catalog is an enterprise


Microsoft
Hortonworks
Power BI
jupyter
Apache Zeppelin
REVOLUTION ANALYTICS
ubuntu
Supported by Canonical
atscale
Datameer
Powerfully Simple
Microsoft Azure Data Lake
Analytics Service
HDInsight
DATAGUISE
cloudera
U-SQL
Spark
STORM
TRIFACTA
SAP Lumira
WATERLINE DATA
tableau
SOFTWARE
MAPR
YARN
HDFS
ZOOMDATA
BlueTalon
Store
platfora
You should know.
CONCURRENT
QlikView
DATATORRENT

Demo Azure Data Lake

Microsoft

If you would like Azure Data Lake Preview Access

Name	Azure Email Account	Azure SubscriptionID
Darwin Schweitzer	datacommoner@outlook.com	bcb1d5d2-e6ea-492d-b9c7-xxxxxxxxxxxx

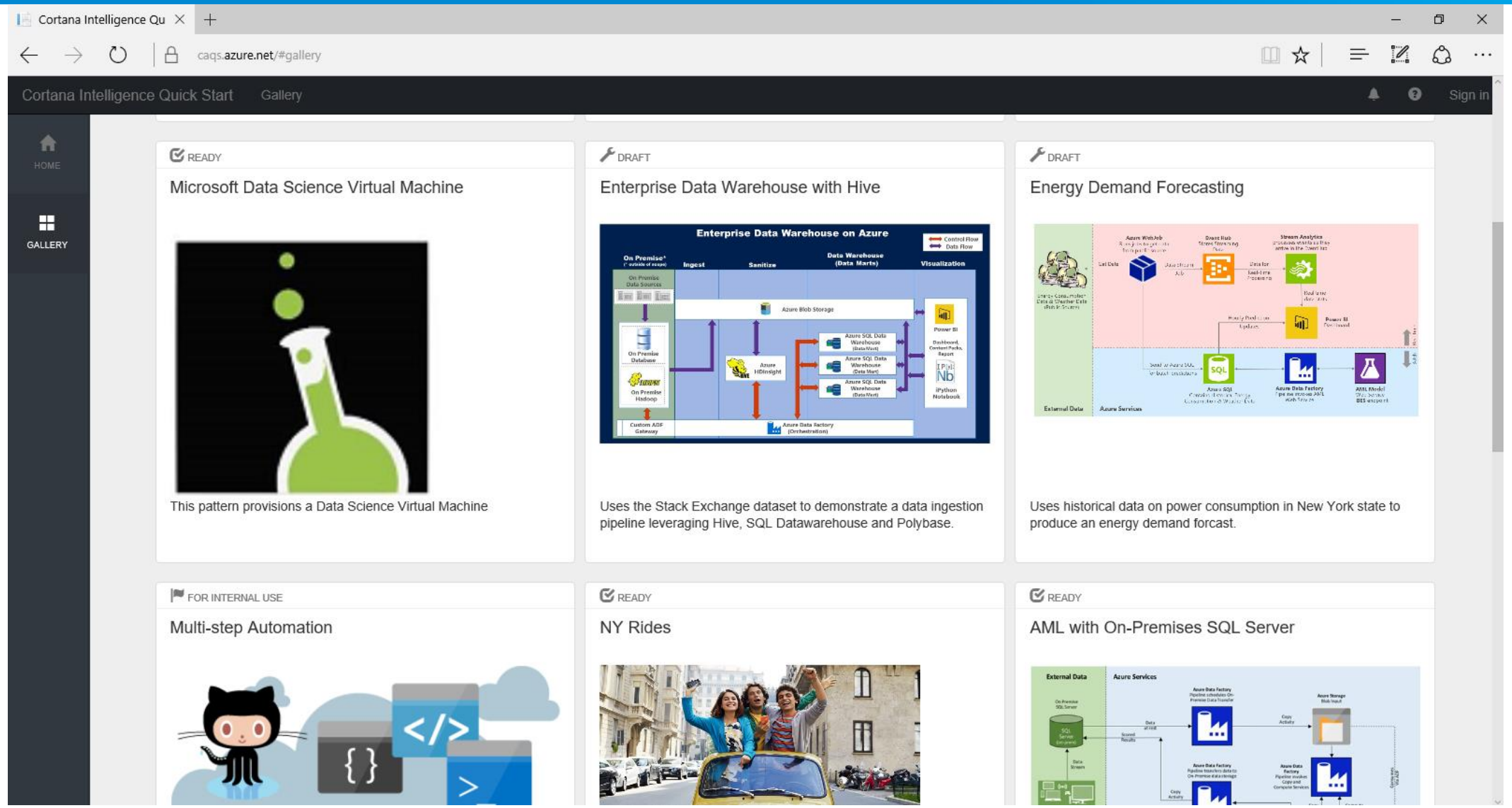
Send email to darsch@microsoft.com

With your:

Name
Azure Email Account
Azure SubscriptionID

To use in HOL or during the Hackathon

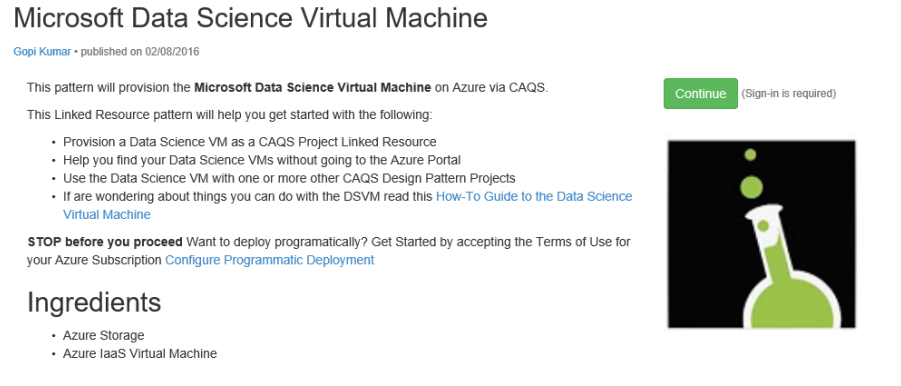
https://caqs.azure.net

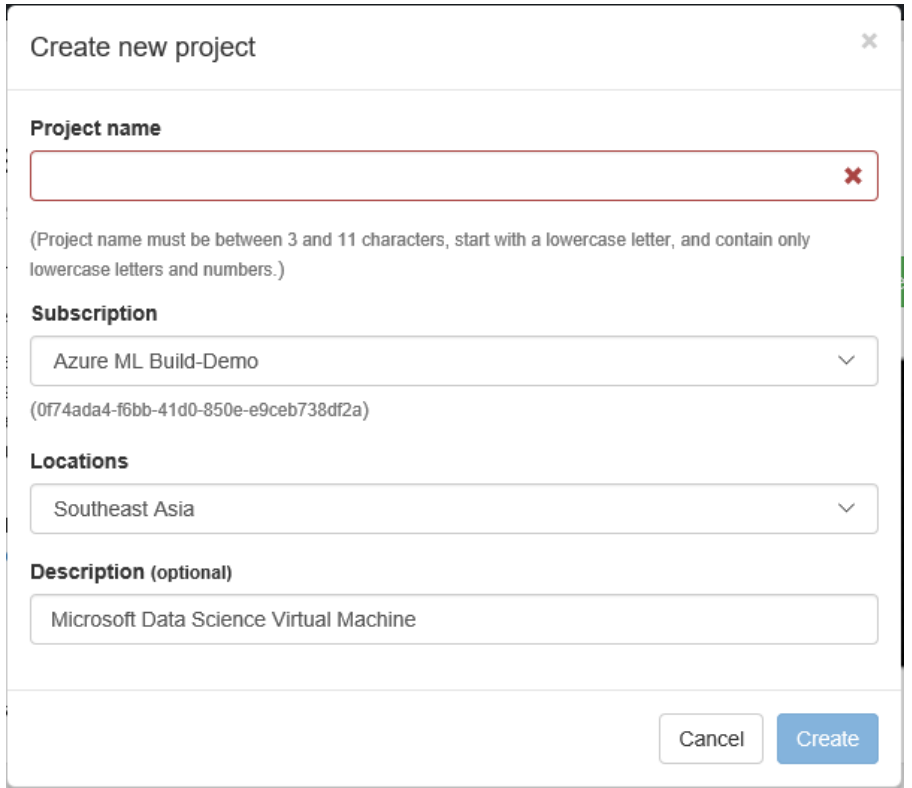


For tomorrows session Power BI with Big Data Stores

Homework

- Go to https://caqs.azure.net/#gallery/datasciencevm
- Sign In with your Azure Subscription account
- Accept the Terms of Use for your Azure Subscription
  Configure Programmatic Deployment
- Click the Continue button to provision
  the Data Science VM
- Fill in parameters and click Create
- Connect to the VM and login

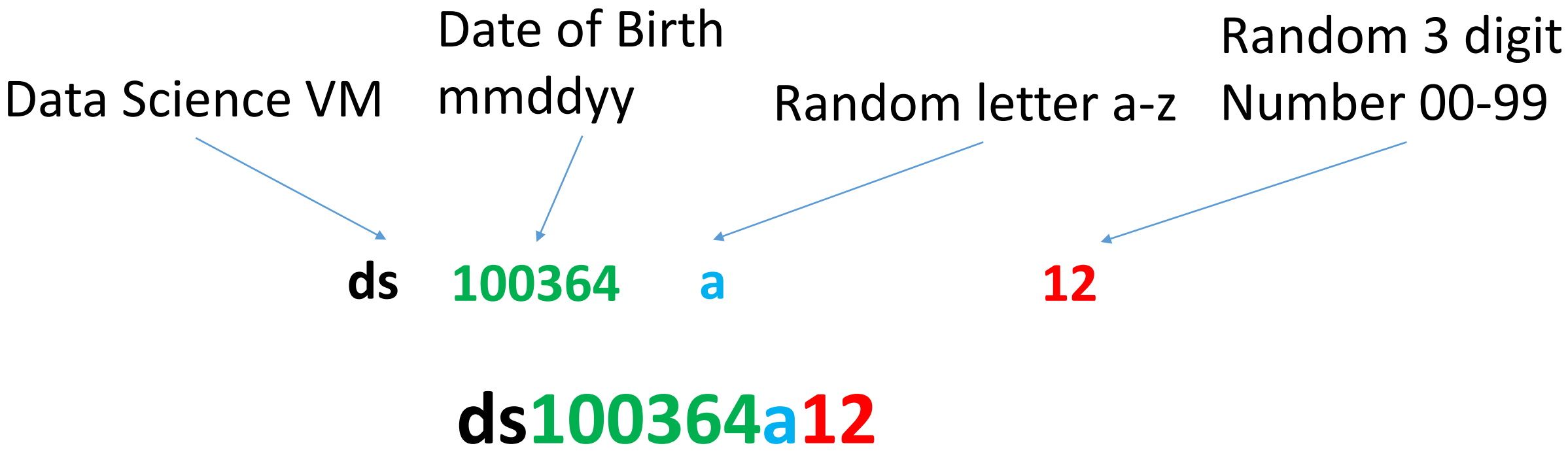


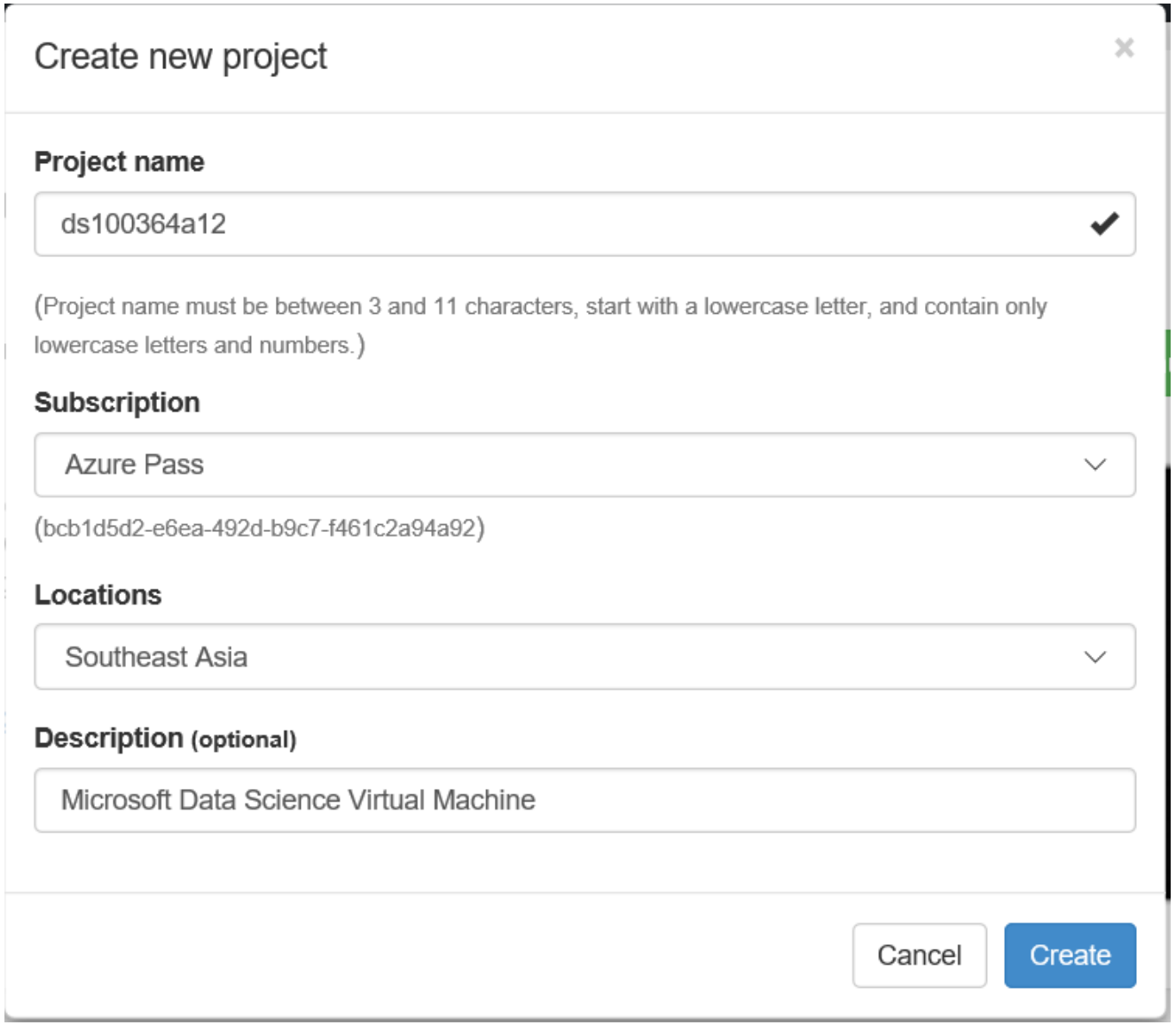


How-To Guide to the Data Science Virtual Machine

CAQS Project Naming recommendation

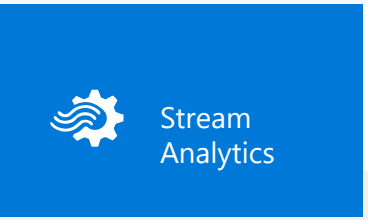
Pattern Id(first two digits) your DOB (next 6 digits mmddyy) Random letter a-z , Random 2 digit number between 00-99

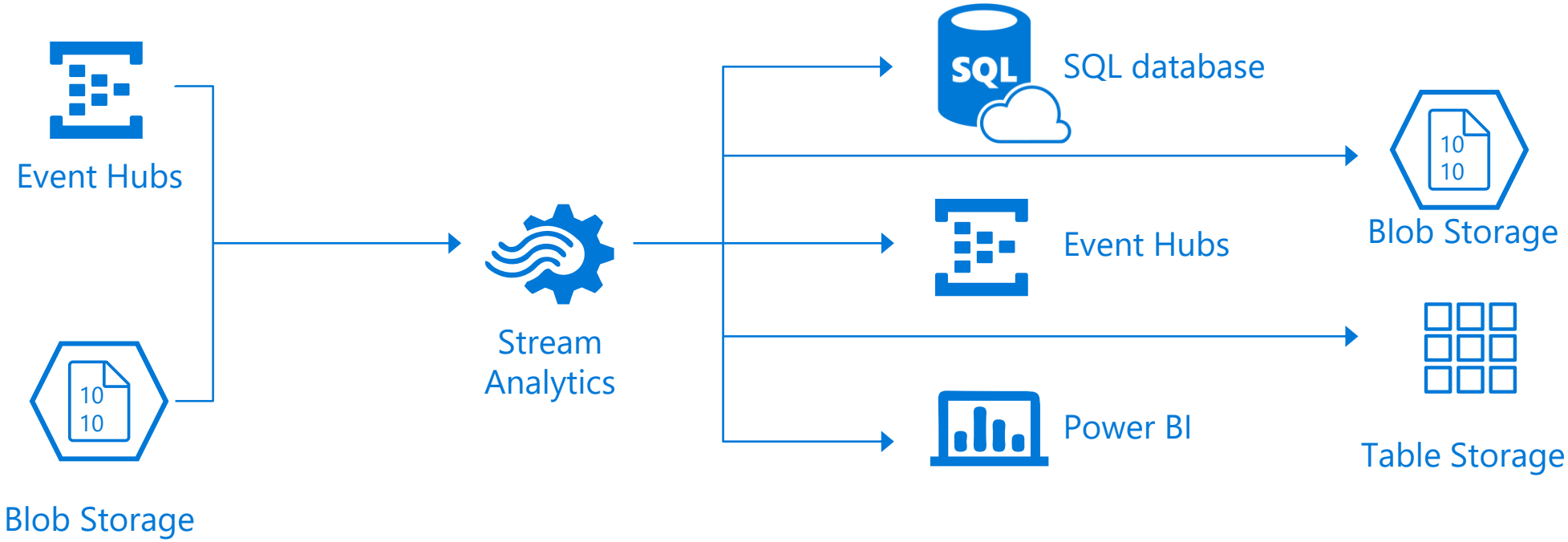



Create new project
Project name
ds100364a12
(Project name must be between 3 and 11 characters, start with a lowercase letter, and contain only lowercase letters and numbers.)
Subscription
Azure Pass
(bcb1d5d2-e6ea-492d-b9c7-f461c2a94a92)
Locations
Southeast Asia
Description (optional)
Microsoft Data Science Virtual Machine
Cancel
Create

Real-time stream processing in the cloud

Darwin Schweitzer
@DataSnowman



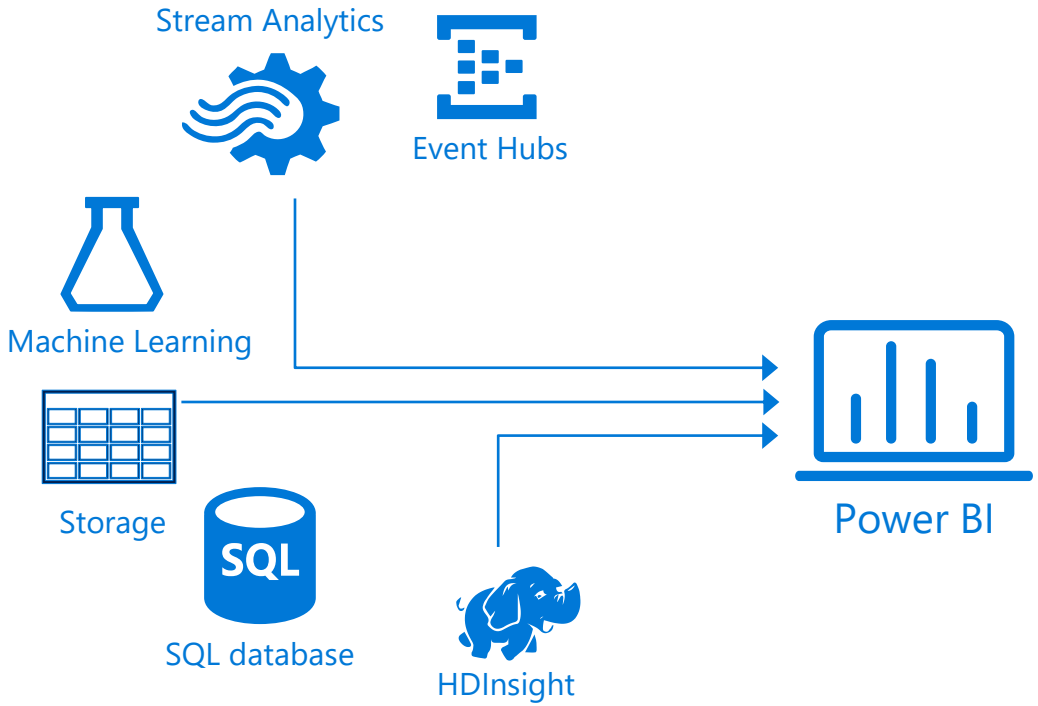


- Perform real-time analytics for your Internet of Things solutions
- Stream millions of events per second
- Get mission-critical reliability and performance with predictable results
- Create real-time dashboards and alerts over data from devices and applications
- Correlate across multiple streams of data
- Use familiar SQL-based language for rapid development

Keep a pulse on your business with live, interactive dashboards

Dashboards & Visualizations

Power BI



- Analytics for everyone, even non-data experts
- Your whole business on one dashboard
- Create stunning, interactive reports
- Drive consistent analysis across your organization
- Embed visuals in your applications
- Get real-time alerts when things change

Power BI – Business Users and Data Analysts

Business analyst tools

Excel

Power BI Desktop

On-premises data

SQL Server analysis services

Databases and other data sources

Microsoft cloud

Power BI

Azure data services

Office 365

Dynamics

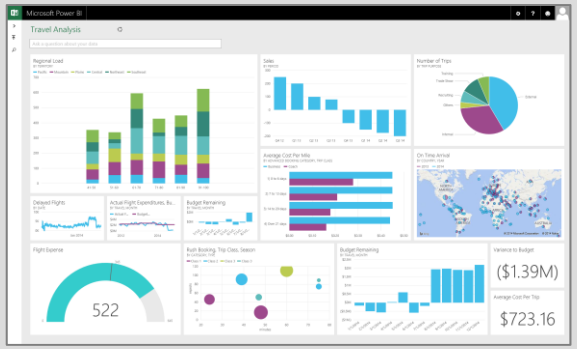
Partner apps

Non-Microsoft cloud

Non Microsoft cloud software systems
(e.g., SaaS applications, IOT scenarios, and data streams)

Web browser

HTML



Mobile apps

Windows

iOS

android

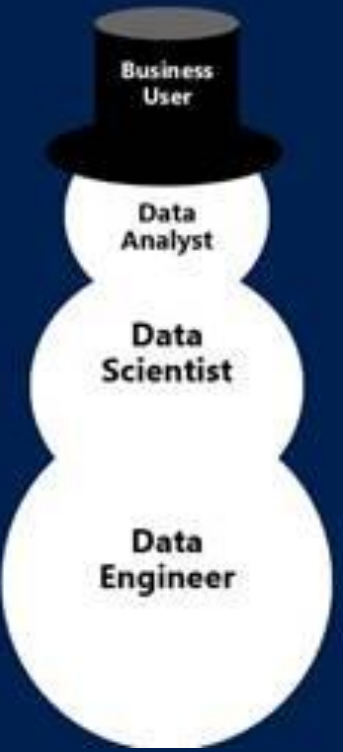

Data
Analyst


Data
Scientist


Start
Data
Engineer

Intuitive
Accessible
- Simple
- Visual

DataSnowman



DASHBOARD MONITOR INPUTS FUNCTIONS PREVIEW QUERY OUTPUTS SCALE CONFIGURE

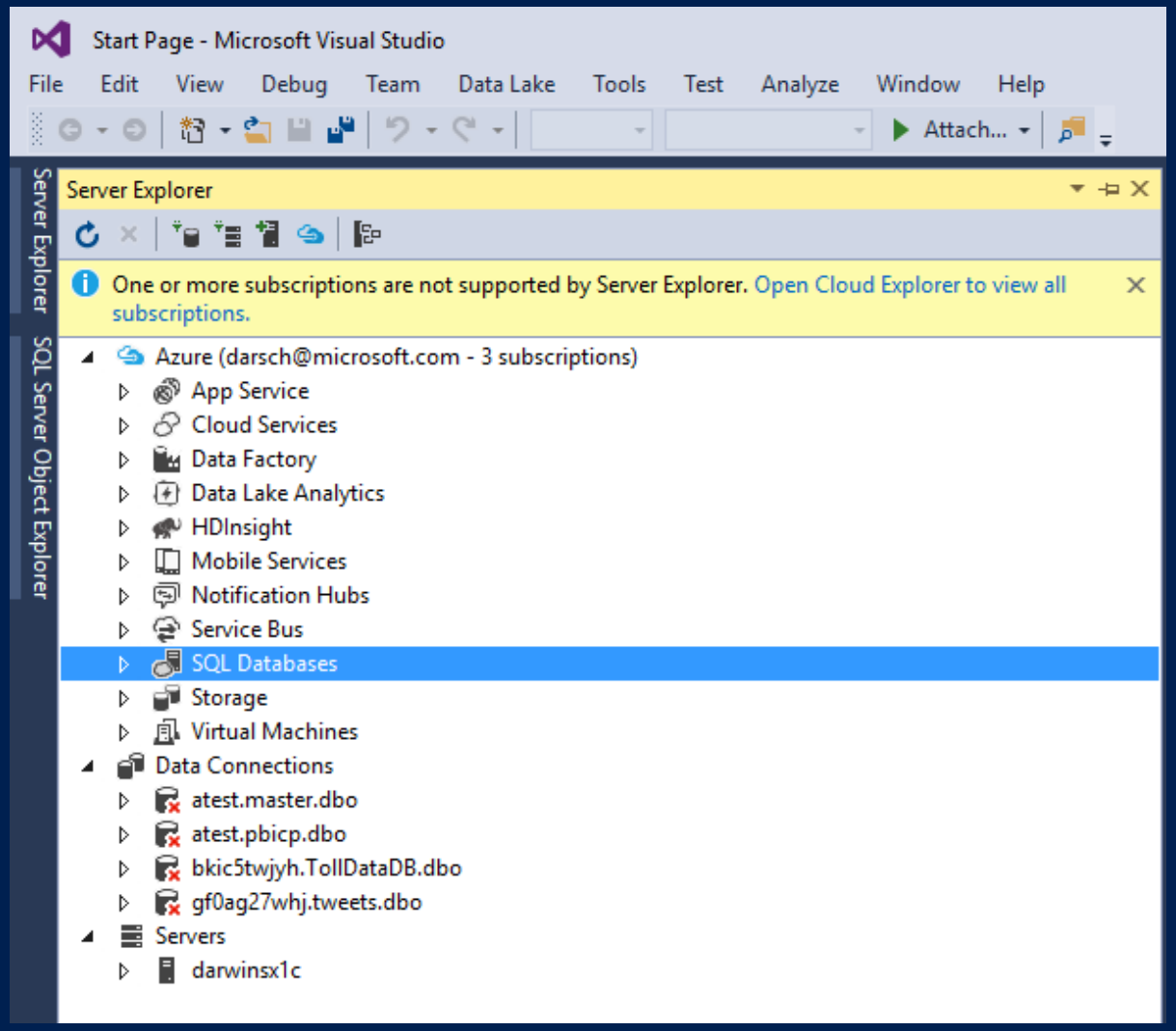
Need help with your query? Check out some of the most common Stream Analytics query patterns here.

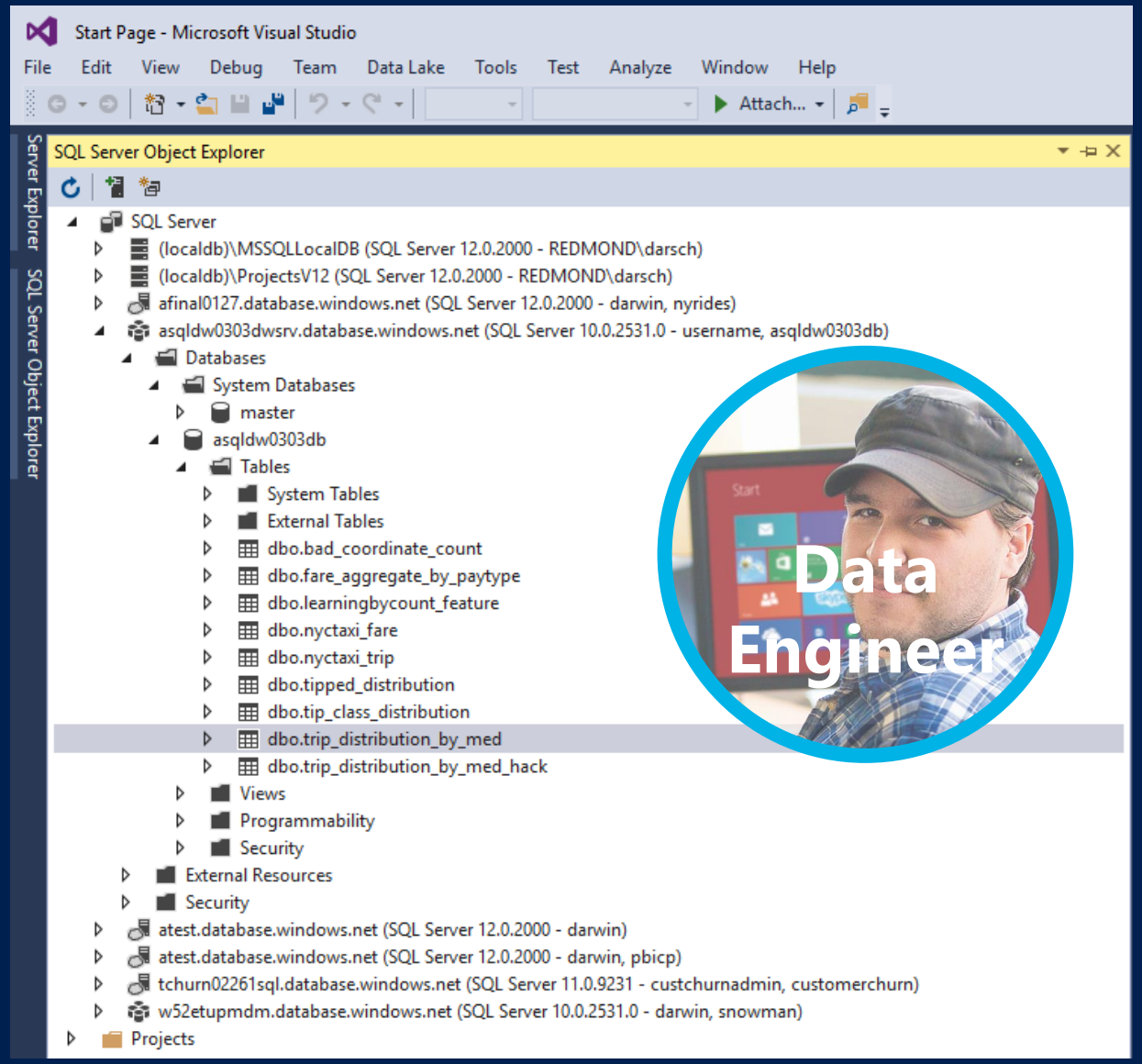
query

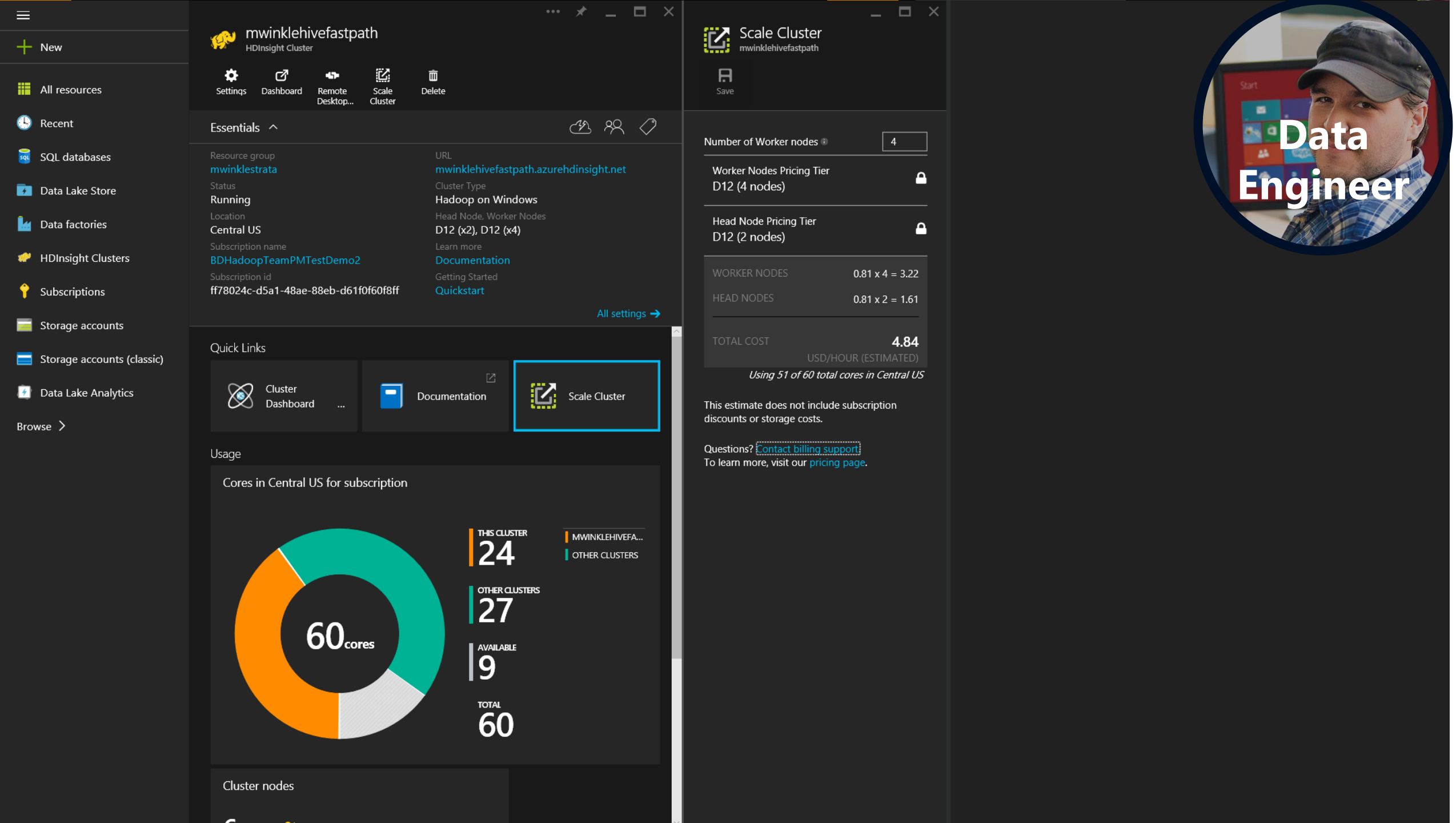
```
1 WITH TipStream As (SELECT System.TimeStamp AS BookingTime, PickupLocation, DestinationLocation,
2 count(*) as NumBookings, max(TipAmount) as MaxTipAmount, min(TipAmount) as MinTipAmount,
3 avg(TipAmount) as AvgTipAmount
4 FROM  bookingstream TIMESTAMP BY CreatedTime
5 GROUP BY PickupLocation, DestinationLocation, TumblingWindow(second,30) )
6 SELECT BookingTime, PickupLocation, DestinationLocation, NumBookings, AvgTipAmount
7 INTO driverstip FROM TipStream
8 SELECT  * INTO  pbipassenger FROM [passengerstream] SELECT * INTO pbidriver FROM [driverstream]
```

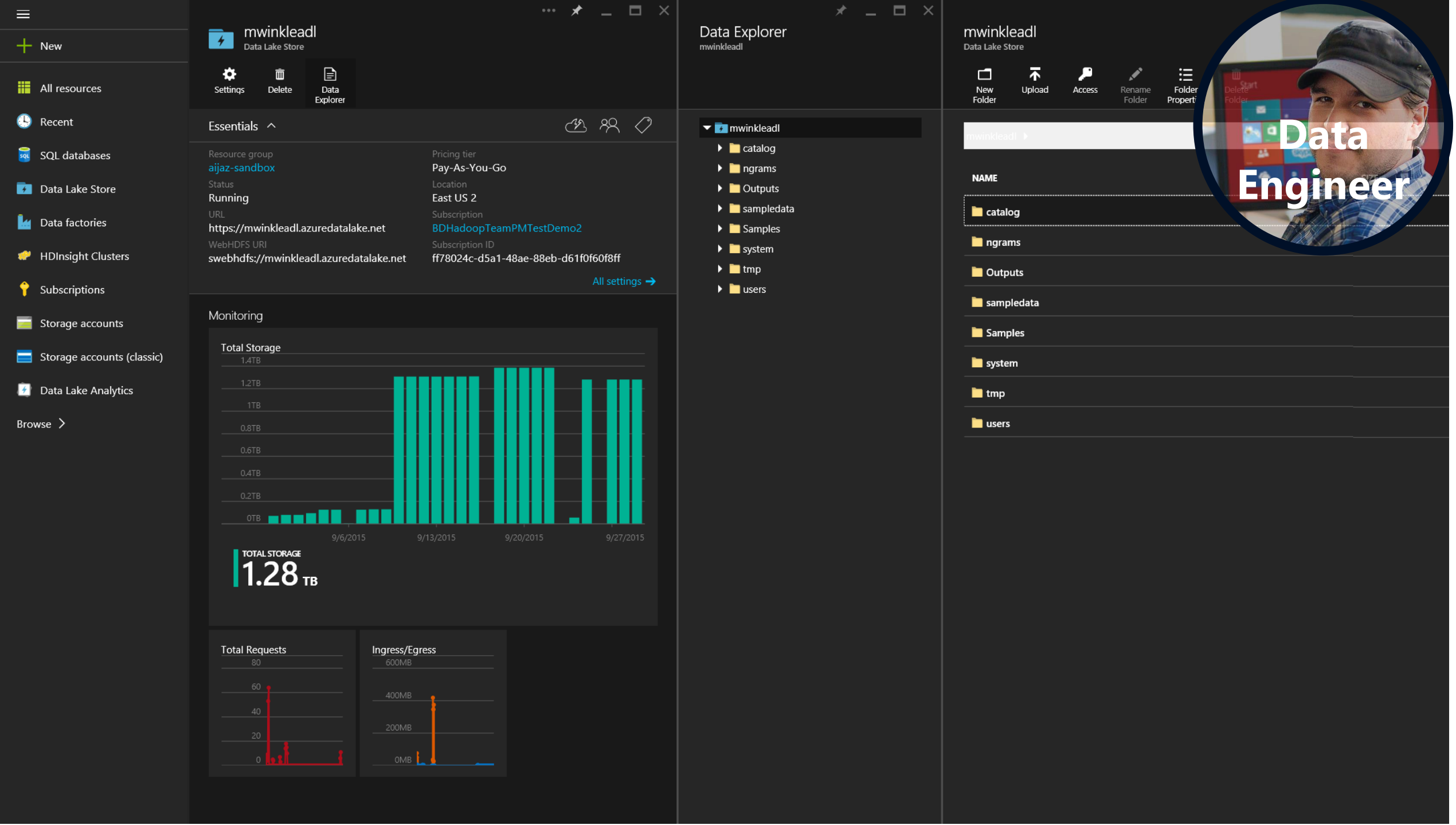
Azure Stream Analytics: easy to write SQL queries over streaming data

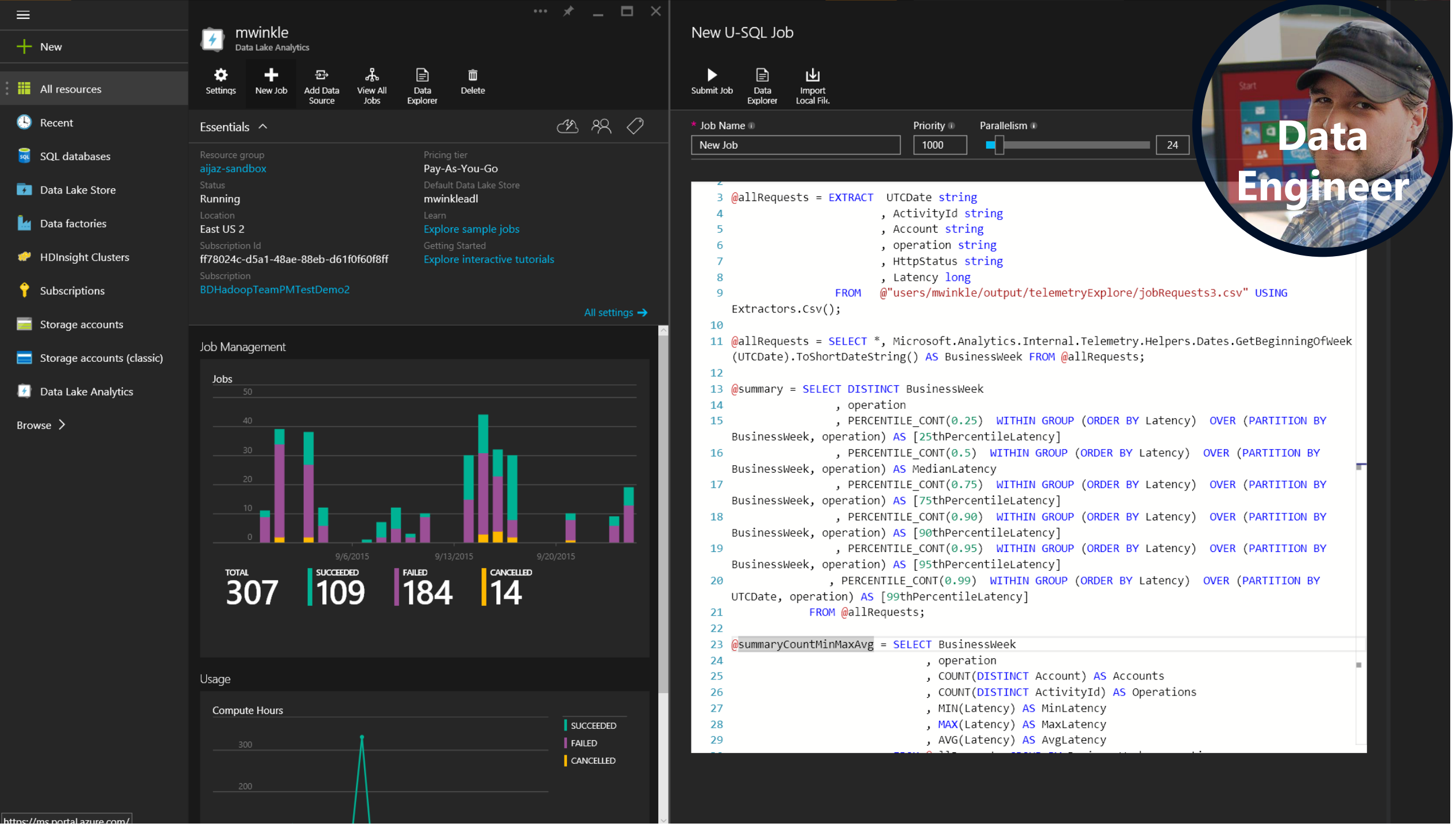
Azure Portal and Visual Studio - Data Engineers






New
All resources
Recent
SQL databases
Data Lake Store
Data factories
HDInsight Clusters
Subscriptions
Storage accounts
Storage accounts (classic)
Data Lake Analytics
Browse
mwinklehivefastpath
HDInsight Cluster
Settings
Dashboard
Remote Desktop...
Scale Cluster
Delete
Essentials
Resource group
mwinklestrata
Status
Running
Location
Central US
Subscription name
BDHadoopTeamPMTestDemo2
Subscription id
ff78024c-d5a1-48ae-88eb-d61f0f60f8ff
URL
mwinklehivefastpath.azurehdinsight.net
Cluster Type
Hadoop on Windows
Head Node, Worker Nodes
D12 (x2), D12 (x4)
Learn more
Documentation
Getting Started
Quickstart
All settings
Quick Links
Cluster Dashboard
Documentation
Scale Cluster
Usage
Cores in Central US for subscription
60 cores
THIS CLUSTER
24
OTHER CLUSTERS
27
AVAILABLE
9
TOTAL
60
MWINKLEHIVEFA...
OTHER CLUSTERS
Cluster nodes
Scale Cluster
mwinklehivefastpath
Save
Number of Worker nodes
4
Worker Nodes Pricing Tier
D12 (4 nodes)
Head Node Pricing Tier
D12 (2 nodes)
WORKER NODES
0.81 x 4 = 3.22
HEAD NODES
0.81 x 2 = 1.61
TOTAL COST
4.84
USD/HOUR (ESTIMATED)
Using 51 of 60 total cores in Central US
This estimate does not include subscription discounts or storage costs.
Questions? Contact billing support
To learn more, visit our pricing page.
Data Engineer


New
All resources
Recent
SQL databases
Data Lake Store
Data factories
HDInsight Clusters
Subscriptions
Storage accounts
Storage accounts (classic)
Data Lake Analytics
Browse
mwinkleadl
Data Lake Store
Settings
Delete
Data Explorer
Essentials
Resource group
aijaz-sandbox
Status
Running
URL
https://mwinkleadl.azuredatalake.net
WebHDFS URI
swebhdfs://mwinkleadl.azuredatalake.net
Pricing tier
Pay-As-You-Go
Location
East US 2
Subscription
BDHadoopTeamPMTestDemo2
Subscription ID
ff78024c-d5a1-48ae-88eb-d61f0f60f8ff
All settings
Monitoring
Total Storage
1.4TB
1.2TB
1TB
0.8TB
0.6TB
0.4TB
0.2TB
0TB
9/6/2015
9/13/2015
9/20/2015
9/27/2015
TOTAL STORAGE
1.28 TB
Total Requests
80
60
40
20
0
Ingress/Egress
600MB
400MB
200MB
0MB
Data Explorer
mwinkleadl
mwinkleadl
catalog
ngrams
Outputs
sampledata
Samples
system
tmp
users
mwinkleadl
Data Lake Store
New Folder
Upload
Access
Rename Folder
Folder Properti
Delete Folder
mwinkleadl
NAME
catalog
ngrams
Outputs
sampledata
Samples
system
tmp
users
Data Engineer


New
All resources
Recent
SQL databases
Data Lake Store
Data factories
HDInsight Clusters
Subscriptions
Storage accounts
Storage accounts (classic)
Data Lake Analytics
Browse
mwinkle
Data Lake Analytics
Settings
New Job
Add Data Source
View All Jobs
Data Explorer
Delete
Essentials
Resource group
aijaz-sandbox
Status
Running
Location
East US 2
Subscription Id
ff78024c-d5a1-48ae-88eb-d61f0f60f8ff
Subscription
BDHadoopTeamPMTestDemo2
Pricing tier
Pay-As-You-Go
Default Data Lake Store
mwinkleadl
Learn
Explore sample jobs
Getting Started
Explore interactive tutorials
All settings
Job Management
Jobs
50
40
30
20
10
0
9/6/2015
9/13/2015
9/20/2015
TOTAL
307
SUCCEEDED
109
FAILED
184
CANCELLED
14
Usage
Compute Hours
300
200
SUCCEEDED
FAILED
CANCELLED
New U-SQL Job
Submit Job
Data Explorer
Import Local File
Job Name
New Job
Priority
1000
Parallelism
24
@allRequests = EXTRACT UTCDate string
, ActivityId string
, Account string
, operation string
, HttpStatus string
, Latency long
FROM @"users/mwinkle/output/telemetryExplore/jobRequests3.csv" USING Extractors.Csv();
@allRequests = SELECT *, Microsoft.Analytics.Internal.Telemetry.Helpers.Dates.GetBeginningOfWeek(UTCDate).ToShortDateString() AS BusinessWeek FROM @allRequests;
@summary = SELECT DISTINCT BusinessWeek
, operation
, PERCENTILE_CONT(0.25) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY BusinessWeek, operation) AS [25thPercentileLatency]
, PERCENTILE_CONT(0.5) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY BusinessWeek, operation) AS MedianLatency
, PERCENTILE_CONT(0.75) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY BusinessWeek, operation) AS [75thPercentileLatency]
, PERCENTILE_CONT(0.90) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY BusinessWeek, operation) AS [90thPercentileLatency]
, PERCENTILE_CONT(0.95) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY BusinessWeek, operation) AS [95thPercentileLatency]
, PERCENTILE_CONT(0.99) WITHIN GROUP (ORDER BY Latency) OVER (PARTITION BY UTCDate, operation) AS [99thPercentileLatency]
FROM @allRequests;
@summaryCountMinMaxAvg = SELECT BusinessWeek
, operation
, COUNT(DISTINCT Account) AS Accounts
, COUNT(DISTINCT ActivityId) AS Operations
, MIN(Latency) AS MinLatency
, MAX(Latency) AS MaxLatency
, AVG(Latency) AS AvgLatency
Data Engineer

Azure Data Lake

Currently in Public Preview

A **cloud scale HDFS store** designed for parallel processing workloads

Accessible to all HDFS compliant analytics applications and tools

No limits to scale

- Petabyte files, unlimited account size
- High throughput to increase analytics performance
- Low latency ingestion with immediate access to data

Intelligent data storage

- Multiple tiers of storage
- Dynamic tiering based workload performance requirements
- Data lifecycle management to manage costs

Enterprise grade security

- AAD based access control of files and folders
- Encryption of data at rest

New

All resources

Recent

SQL databases

Data Lake Store

Data factories

HDInsight Clusters

Subscriptions

Storage accounts

Storage accounts (classic)

Data Lake Analytics

Browse

AzureBigAnalyticsDataLakeSample

Data factory

Delete

Essentials

Resource group
ADF

Location
WestUS

Subscription name
MDP_492270

Provisioning state
Succeeded

Subscription id
1e42591f-1f0c-4c5a-b7f2-a268f6105ec5

All settings

Summary

Author and deploy

Diagram

No permission to resource group: ADF

Contents

Datasets 4

With errors 2

Pipelines 3

ComputeEventsByRegion...

EgressBlobToDataLakePi...

...

Linked services 4

Data Stores 4

Data Gateways 0

...

Operations

Events
AZUREBIGANALYTICSDATALAKESAMPLE

Alert rules
AZUREBIGANALYTICSDATALA...

No access

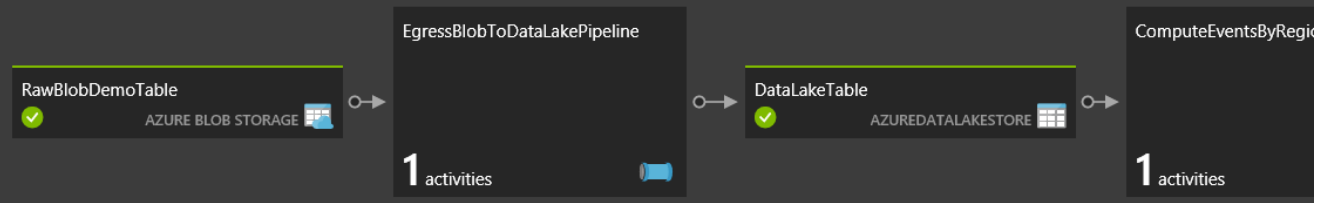
Monitoring

Data factory metrics

Diagram

AzureBigAnalyticsDataLakeSample

Data factory



DevDaysAsia 2016 @ Taipei

Questions or Comments?

Microsoft

DevDaysAsia 2016 @ Taipei

Thank You!

darsch@microsoft.com

Microsoft