Microsoft

Cortana Analytics Workshop

Sept 10 – 11, 2015 • MSCC

Real-Time Data Processing: How Do I Choose the Right Solution?

Benjamin Wright-Jones
Solution Architect, Data Insights Global Practice

Simon Lidberg
Solution Architect, Data Insights Global Practice

Microsoft

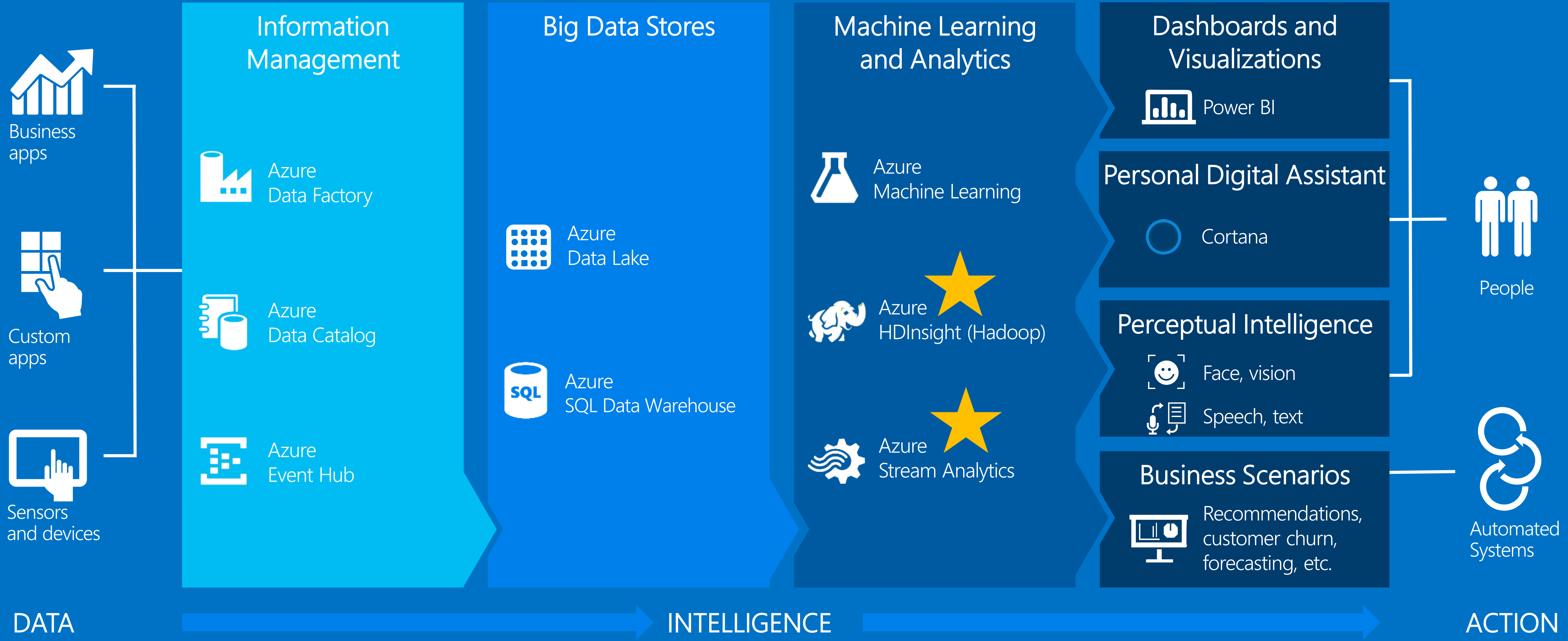
Data Insights Global Practice

A highly specialized team of architects and subject-matter experts

- Customer conversations & workshops
- Solution accelerators & architecture
- Deployment & delivery services
- Knowledge transfer & best practices

Cortana Analytics Suite

Transform data into intelligent action



Business apps

Custom apps

Sensors and devices

Information Management

Azure Data Factory

Azure Data Catalog

Azure Event Hub

Big Data Stores

Azure Data Lake

Azure SQL Data Warehouse

Machine Learning and Analytics

Azure Machine Learning

Azure HDInsight (Hadoop)

Azure Stream Analytics

Dashboards and Visualizations

Power BI

Personal Digital Assistant

Cortana

Perceptual Intelligence

Face, vision

Speech, text

Business Scenarios

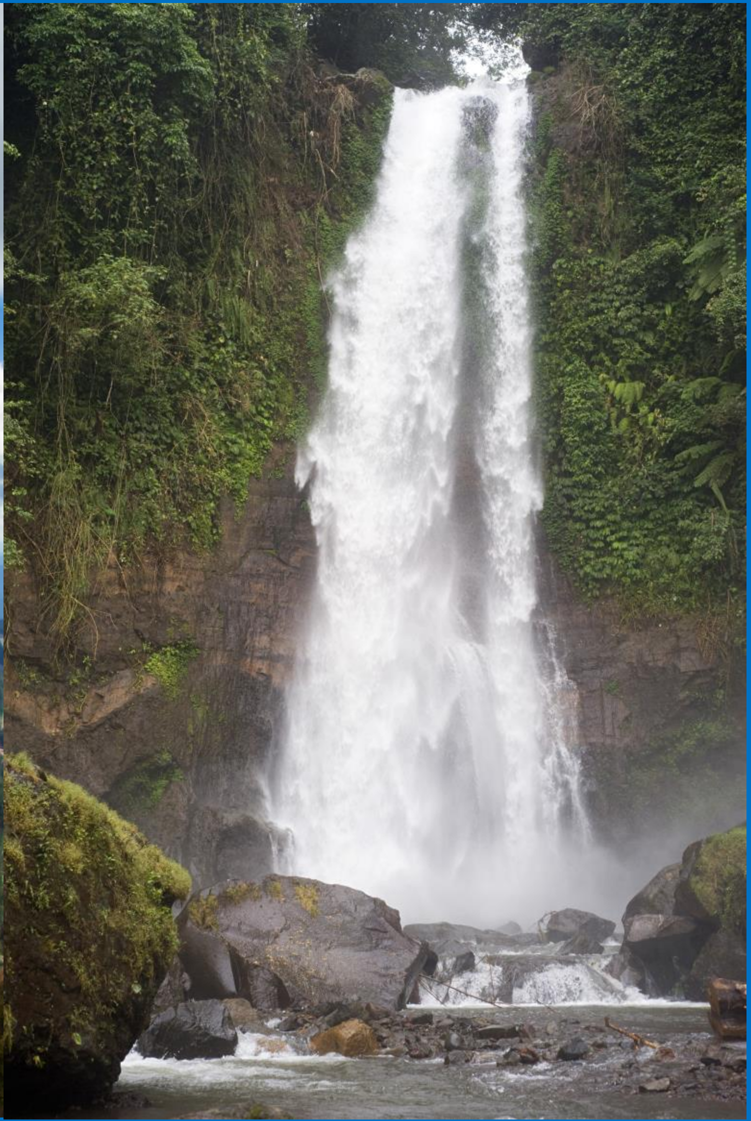
Recommendations, customer churn, forecasting, etc.

People

Automated Systems

DATA → INTELLIGENCE → ACTION

Data Streaming 101



Customers are looking to derive more and more value from data…

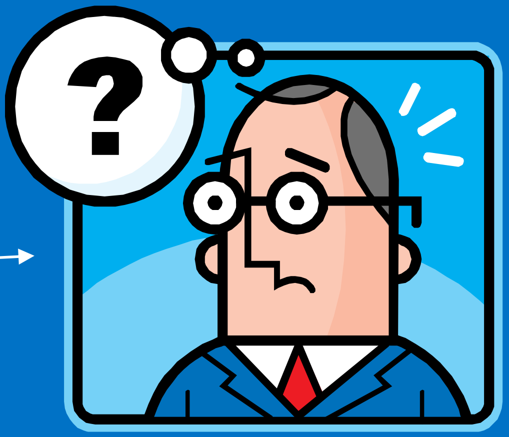
EXAMPLE SOLUTIONS

Sales and marketing	Finance and risk	Customer and channel	Operations and workforce
Customer Acquisition	Fraud detection	Lifetime customer value	Pay for performance
Cross-sell and upsell	Credit risk management	Personalized offers	Operational efficiency
Loyalty programs		Product recommendation	Smart buildings
Marketing mix optimization			Predictive maintenance
			Supply chain management

Increasing number of choices

Three streaming options but which one do I choose?

What are the decision points?



Storm on HDInsight

Azure Stream Analytics

Spark Streaming on HDInsight

Azure Stream Analytics

Fully managed service

No hardware deployment

Scalable

Dynamically scalable

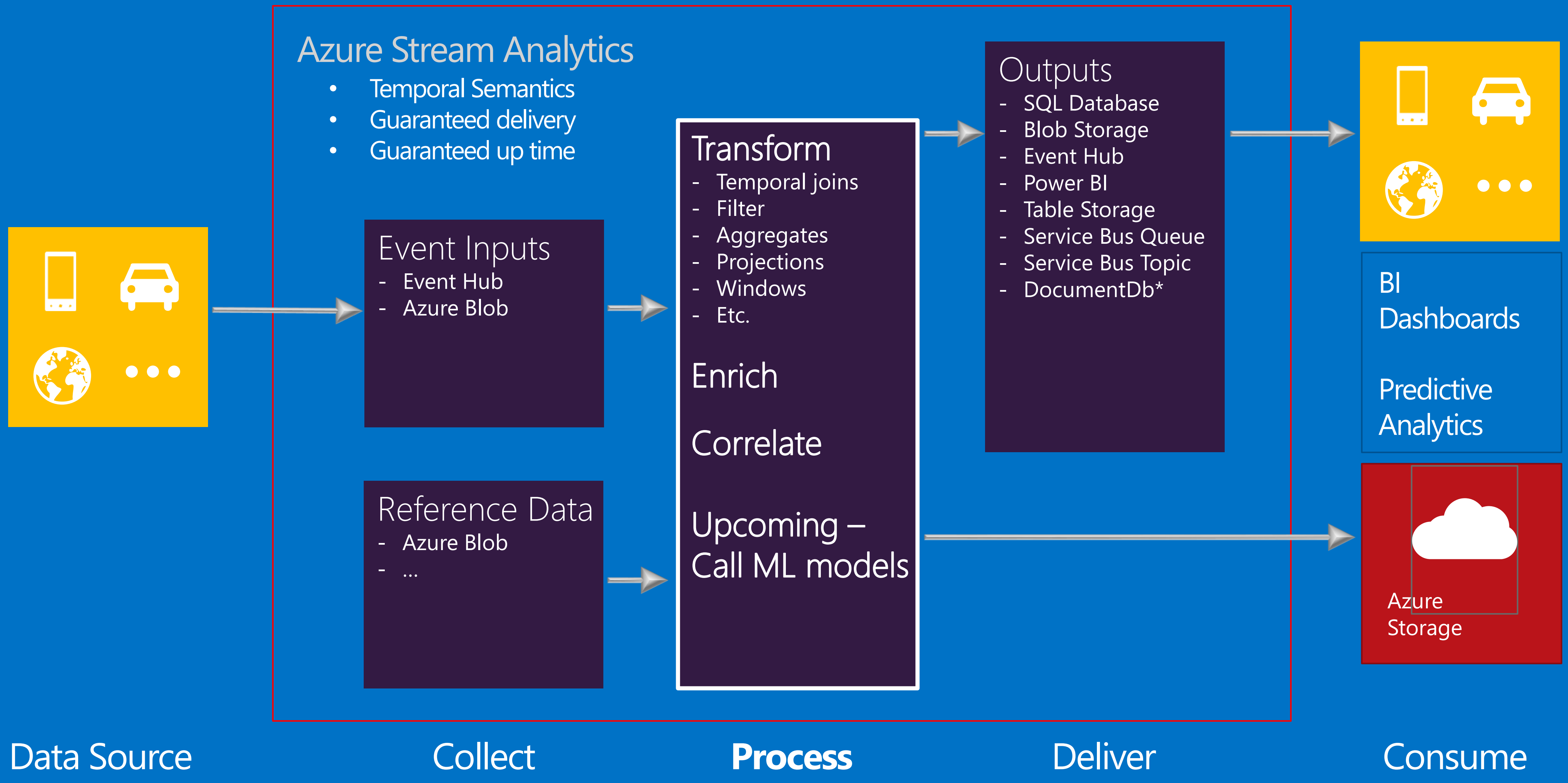
Easy development

SQL Language

Built-in monitoring

View system performance through Azure portal

End-to-End Architecture Overview



Azure Stream Analytics

- Temporal Semantics
- Guaranteed delivery
- Guaranteed up time

Event Inputs
- Event Hub
- Azure Blob

Reference Data
- Azure Blob
- …

Transform
- Temporal joins
- Filter
- Aggregates
- Projections
- Windows
- Etc.

Enrich

Correlate

Upcoming – Call ML models

Outputs
- SQL Database
- Blob Storage
- Event Hub
- Power BI
- Table Storage
- Service Bus Queue
- Service Bus Topic
- DocumentDb*

BI Dashboards

Predictive Analytics

Azure Storage

Data Source

Collect

Process

Deliver

Consume

Demo

Azure Stream Analytics

Query Language - Overview

DML Statements

- SELECT
- INTO
- FROM
- WHERE
- GROUP BY
- HAVING
- CASE
- JOINS
- UNION
- WITH
- CROSS/OUTER APPLY

Scaling Functions

- WITH
- PARTITION BY

Conversion Functions

- CAST

Date and Time Functions

- DATENAME
- DATEPART
- DAY
- MONTH
- YEAR
- DATETIMEFROMPARTS
- DATEDIFF
- DATADD

Windowing Extensions

- Tumbling Window
- Hopping Window
- Sliding Window

Analytic Functions

- ISFIRST
- LAG
- LAST

Aggregate Functions

- SUM
- COUNT
- AVG
- MIN
- MAX
- STDEV
- STDEVP
- VAR
- VARP
- CollectTOP

String Functions

- LEN
- CONCAT
- CHARINDEX
- SUBSTRING
- PATINDEX
- LOWER
- UPPER
- ARRAY

Built in Temporal Semantics

Easily implement temporal functions

Tumbling Windows

Repeating, non-overlapping, fixed interval windows

Hopping Windows

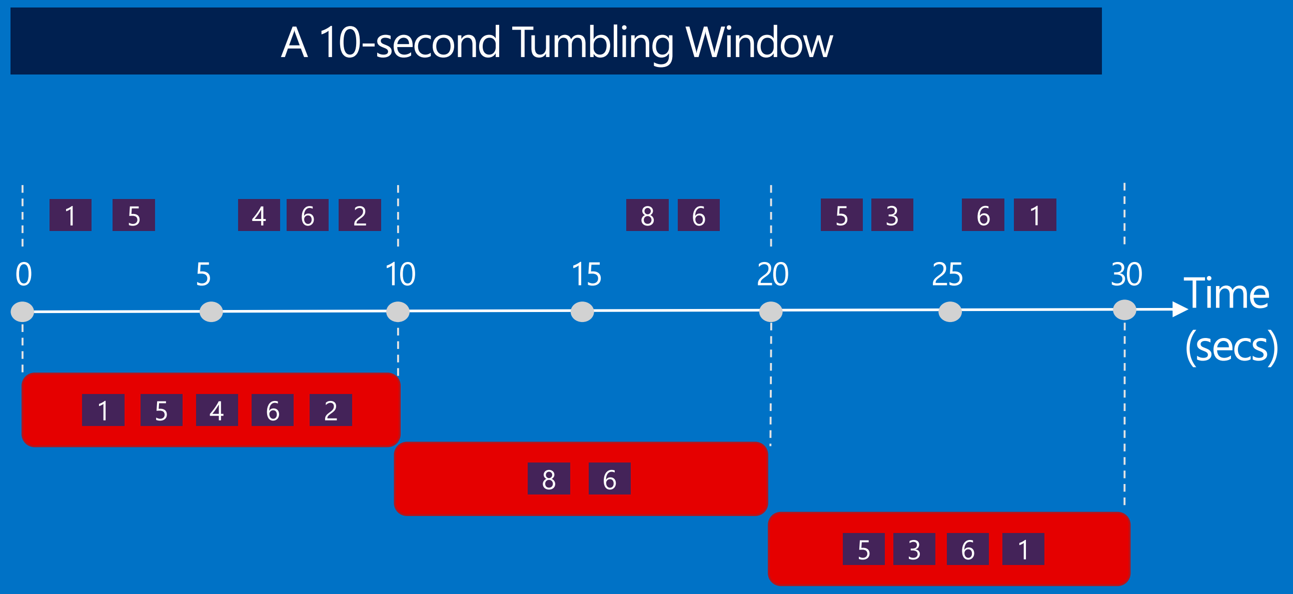
Generic window, overlapping, fixed size

Sliding Windows

Slides by an epsilon and produces output at the occurrence of an event

Tumbling Windows

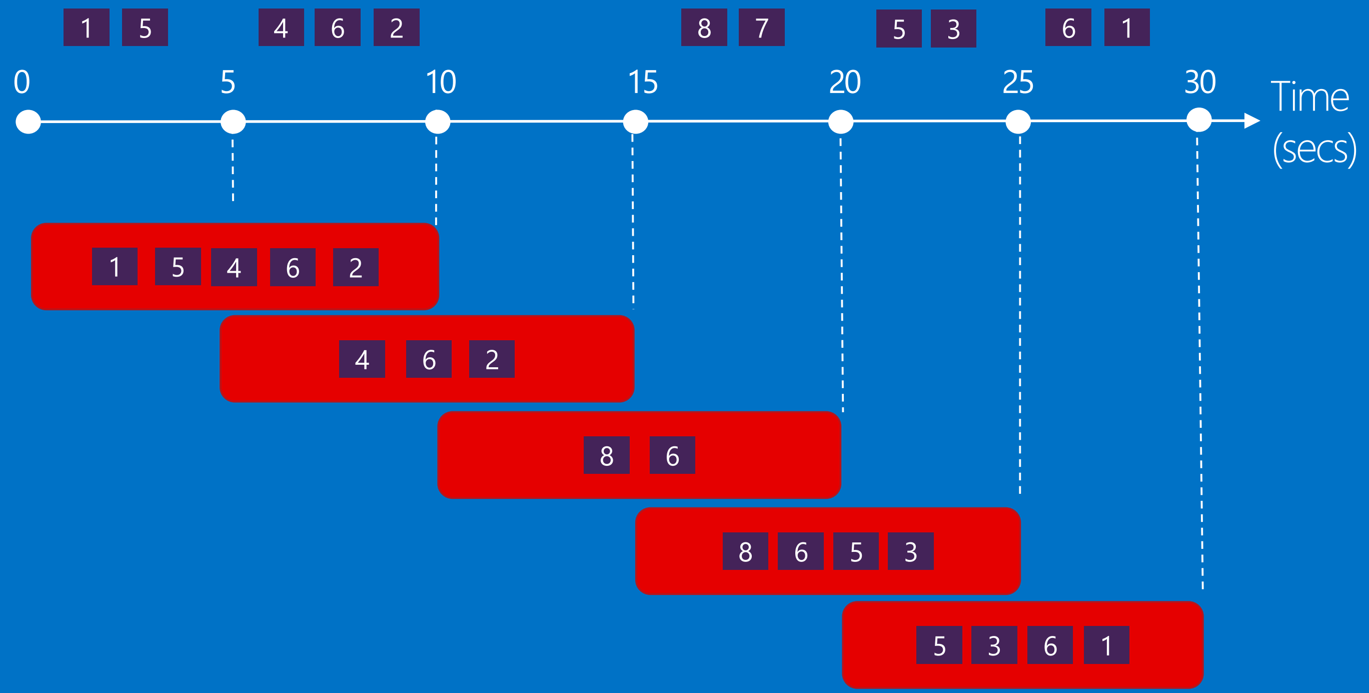
Tell me the count of tweets per time zone every 10 seconds



```
SELECT TimeZone, COUNT(*) AS Count
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY TimeZone, TumblingWindow(second,10)
```

Hopping Windows

A 10-second Hopping Window with a 5-second "Hop"

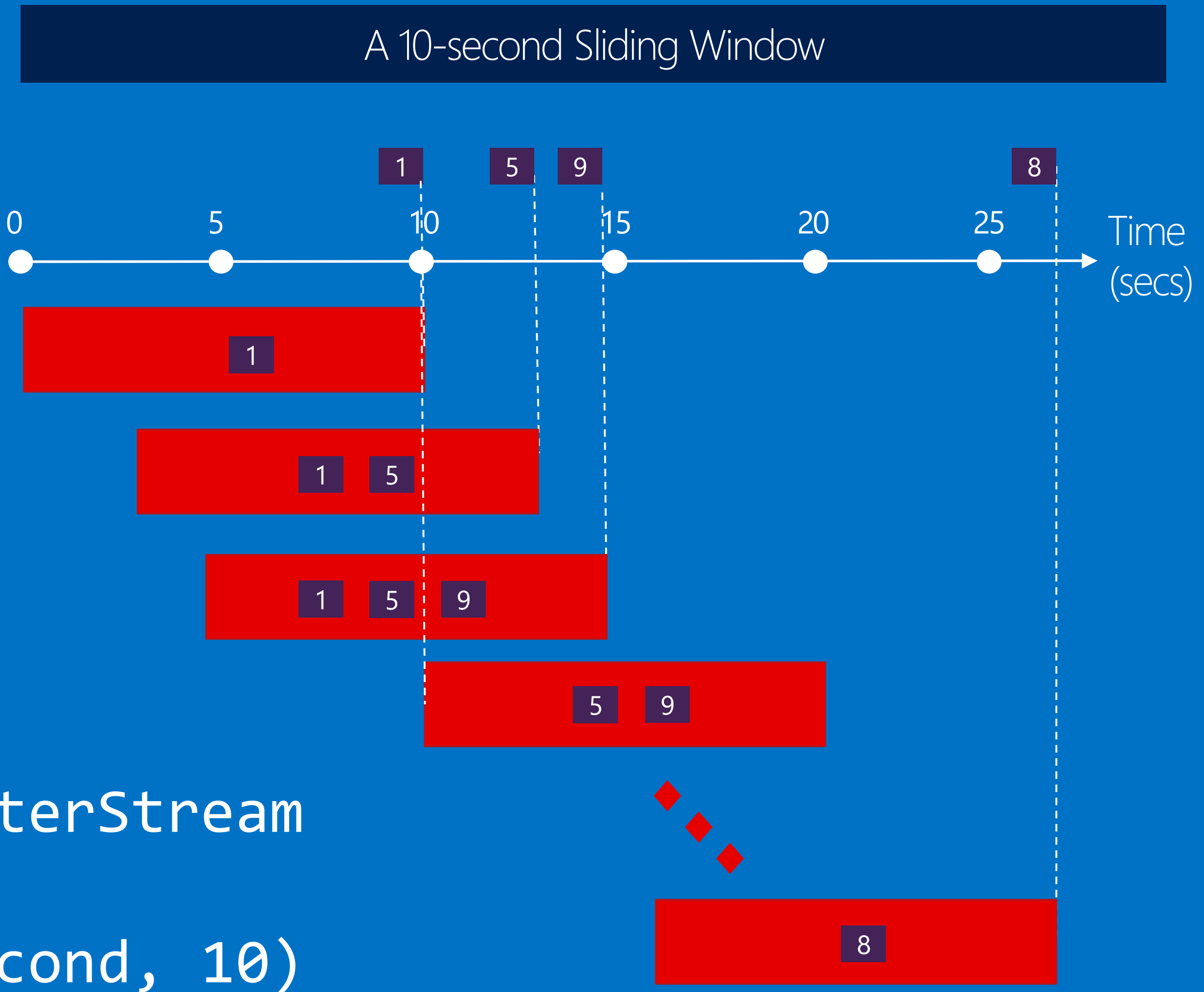


Every 5 seconds give me the count of tweets and the average sentiment score over the last 10 seconds

```
SELECT Topic, COUNT(*) AS TotalTweets, AVG(SentimentScore)
FROM TwitterStream TIMESTAMP BY CreatedAt
GROUP BY Topic, HoppingWindow(second, 10 , 5)
```

Sliding Windows

Give me the count of tweets for all topics which are tweeted more than 10 times in the last 10 seconds



A 10-second Sliding Window

```
SELECT Topic, COUNT(*) FROM TwitterStream
TIMESTAMP BY CreatedAt
GROUP BY Topic, SlidingWindow(second, 10)
HAVING COUNT(*) > 10
```

Business transformation

Asthma device manufacturer uses the cloud to improve data collection and in the process reshapes its business towards greater efficiency

Aerocrine

Objectives

- Aerocrine produces devices that help monitor asthma for sufferers. The devices are sensitive to small changes in the ambient environment. It wanted to improve the effectiveness of the devices.

Tactics

- Implemented Azure to gather device data
- Developed an application to transmit data
- Used Azure Events Hub and Azure Stream Analytics for analysis

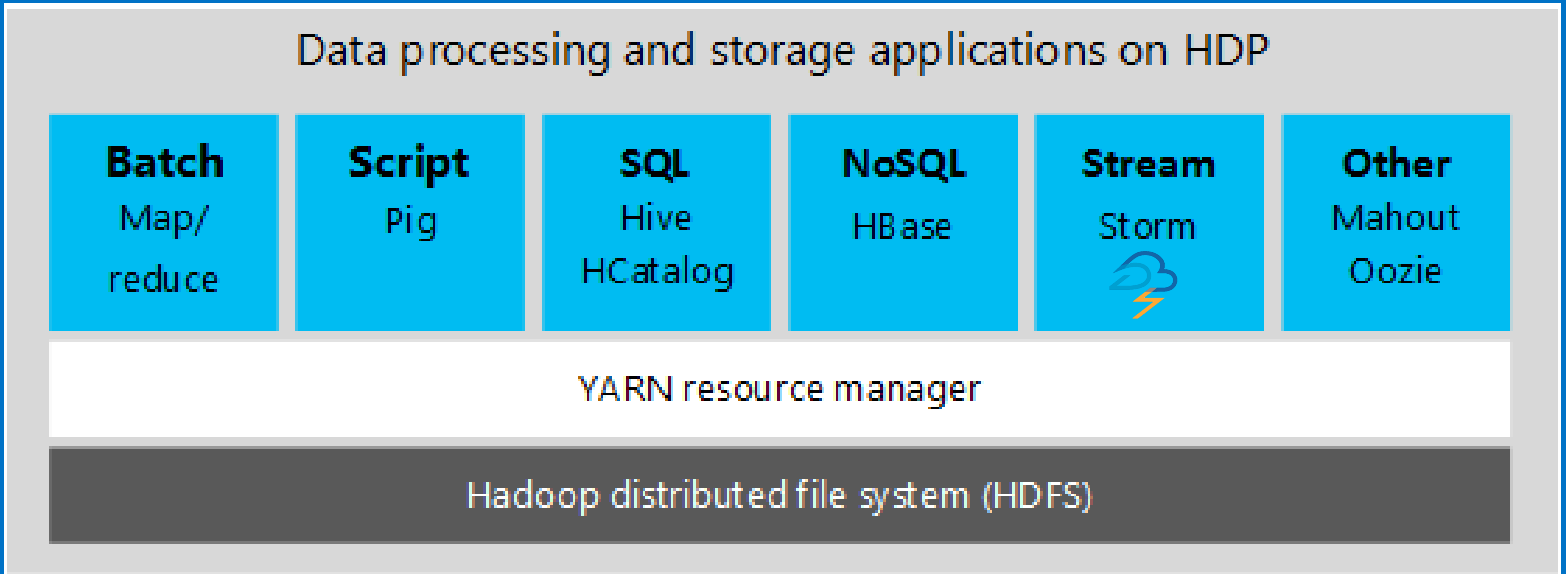
Results

- Collects near real-time telemetry data
- Discovers trigger points that affect devices
- Transforming business model with greater insight into device operation
- Ultimately aims to help physicians provide better levels of care

> “With the Microsoft Azure solution, we are getting much deeper analysis into our devices. That means we can better identify the trigger points that are affecting device performance.”
>
> —Anders Murman, Chief Technology Officer, Aerocrine

Storm on HDInsight

What is Storm?



Data processing and storage applications on HDP

Batch Map/ reduce

Script Pig

SQL Hive HCatalog

NoSQL HBase

Stream Storm

Other Mahout Oozie

YARN resource manager

Hadoop distributed file system (HDFS)

Built on hadoop

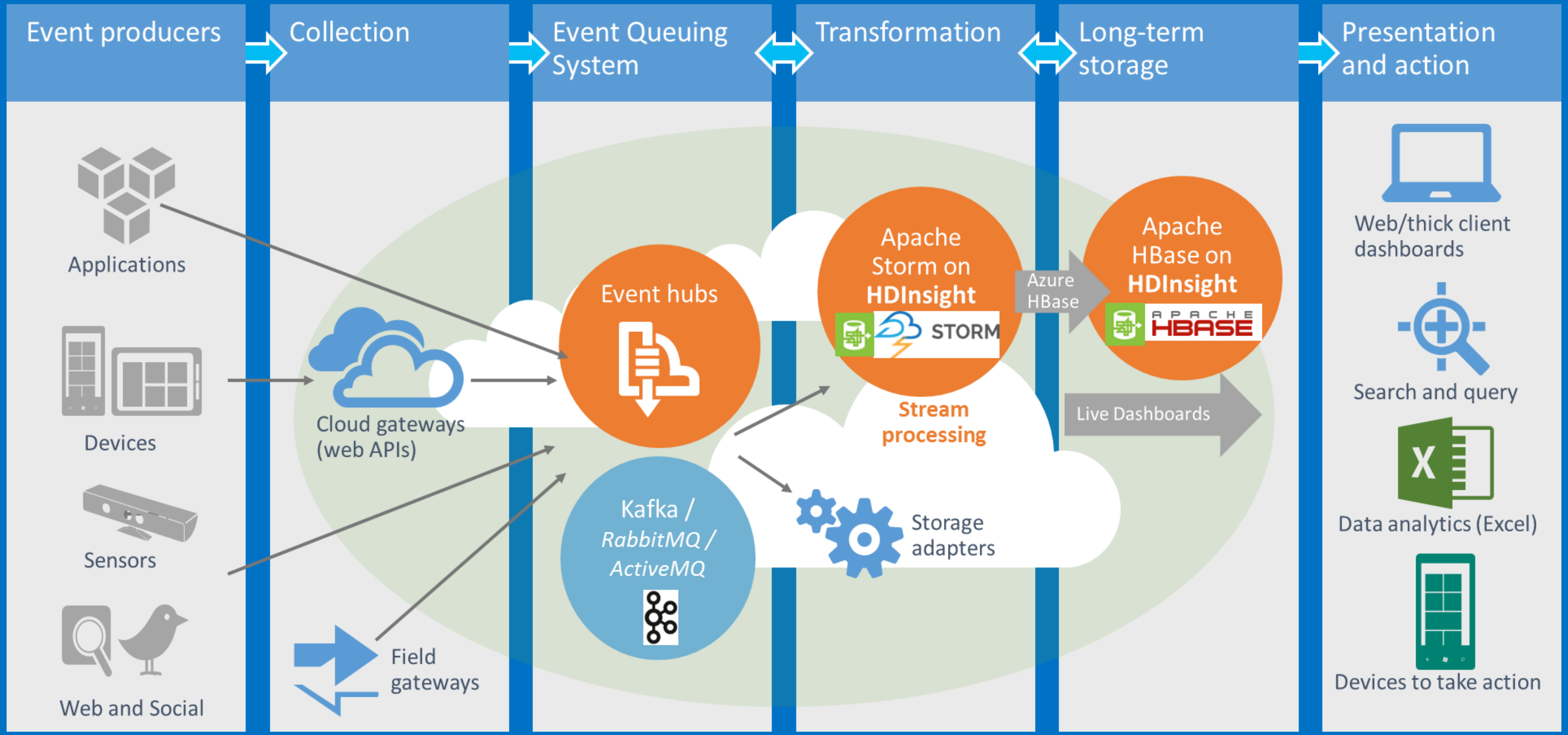
What is Storm?

Real-time Stream Processing

Open Source

Visual Studio Integration

Available on Azure HDInsight

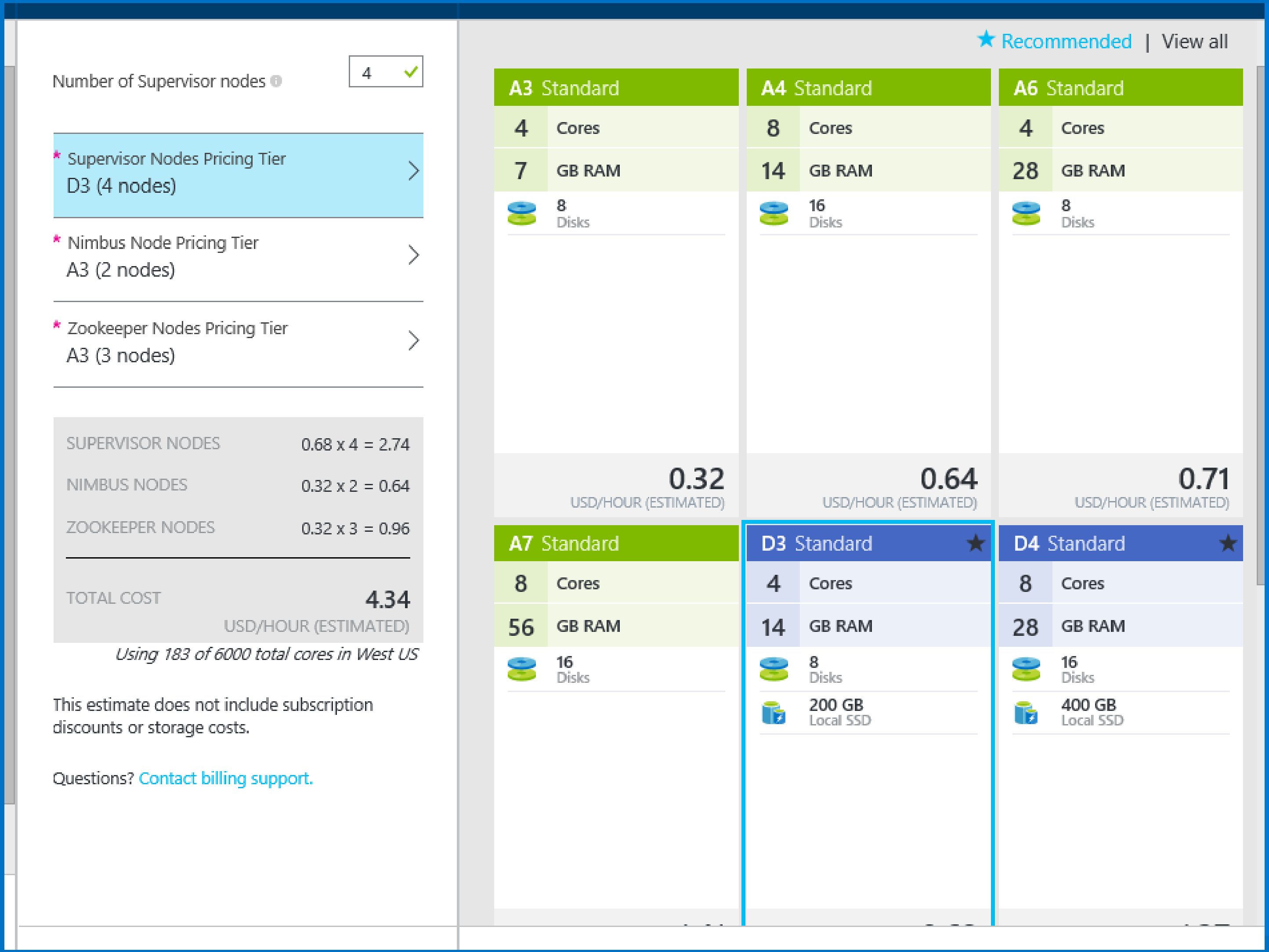


What is in a HDInsight Storm Cluster?

Flexible choice

Dynamic Rebalance

1-n nodes*



Number of Supervisor nodes: 4 ✓

* Supervisor Nodes Pricing Tier
D3 (4 nodes)

* Nimbus Node Pricing Tier
A3 (2 nodes)

* Zookeeper Nodes Pricing Tier
A3 (3 nodes)

SUPERVISOR NODES	0.68 x 4 = 2.74
NIMBUS NODES	0.32 x 2 = 0.64
ZOOKEEPER NODES	0.32 x 3 = 0.96
TOTAL COST	4.34 USD/HOUR (ESTIMATED)

Using 183 of 6000 total cores in West US

This estimate does not include subscription discounts or storage costs.

Questions? Contact billing support.

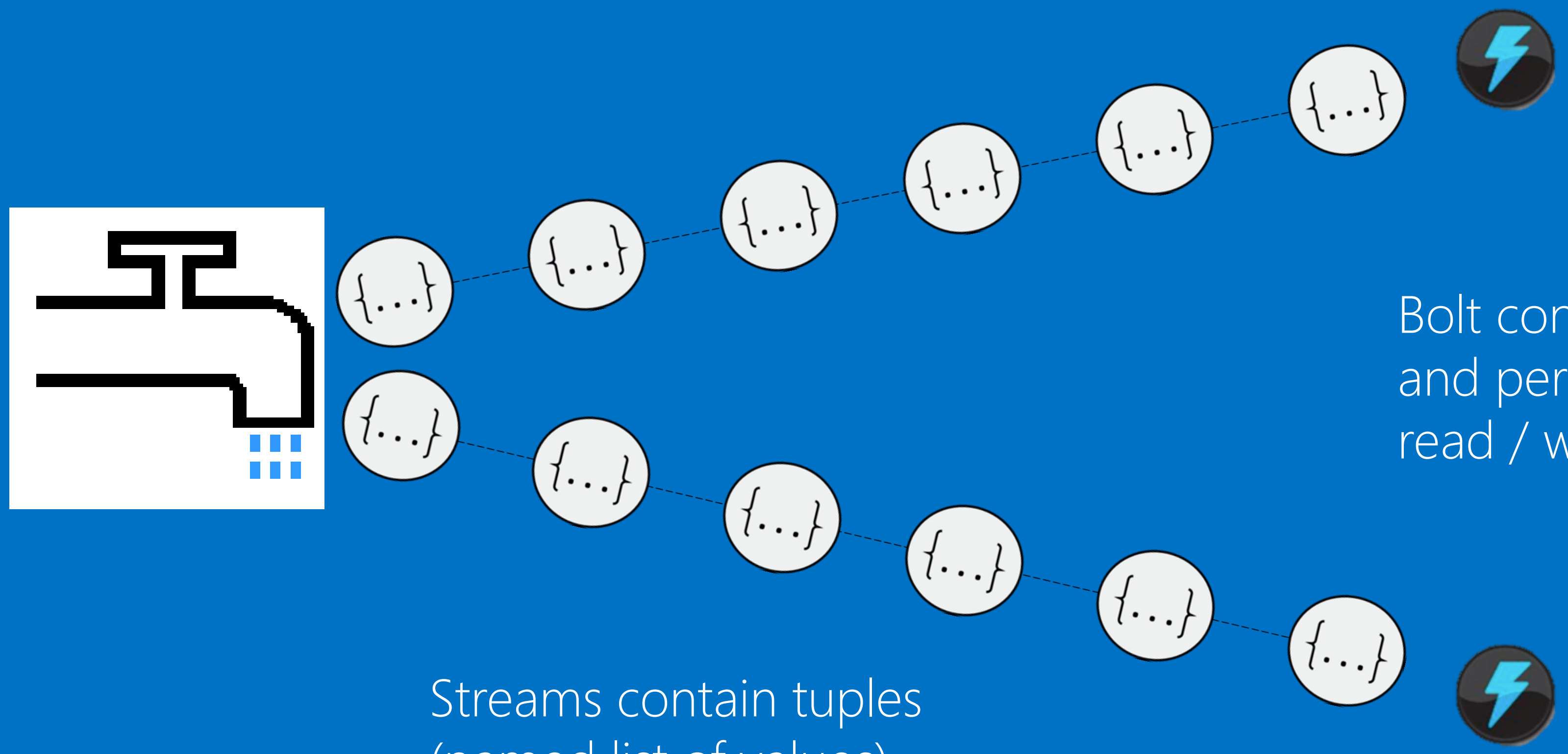
★ Recommended | View all

Tier	Cores	GB RAM	Disks	Local SSD	USD/HOUR (ESTIMATED)
A3 Standard	4	7	8		0.32
A4 Standard	8	14	16		0.64
A6 Standard	4	28	8		0.71
A7 Standard	8	56	16		
D3 Standard ★	4	14	8	200 GB	
D4 Standard ★	8	28	16	400 GB	

http://azure.microsoft.com/en-us/pricing/details/hdinsight/

Spouts, Bolts and Tuples

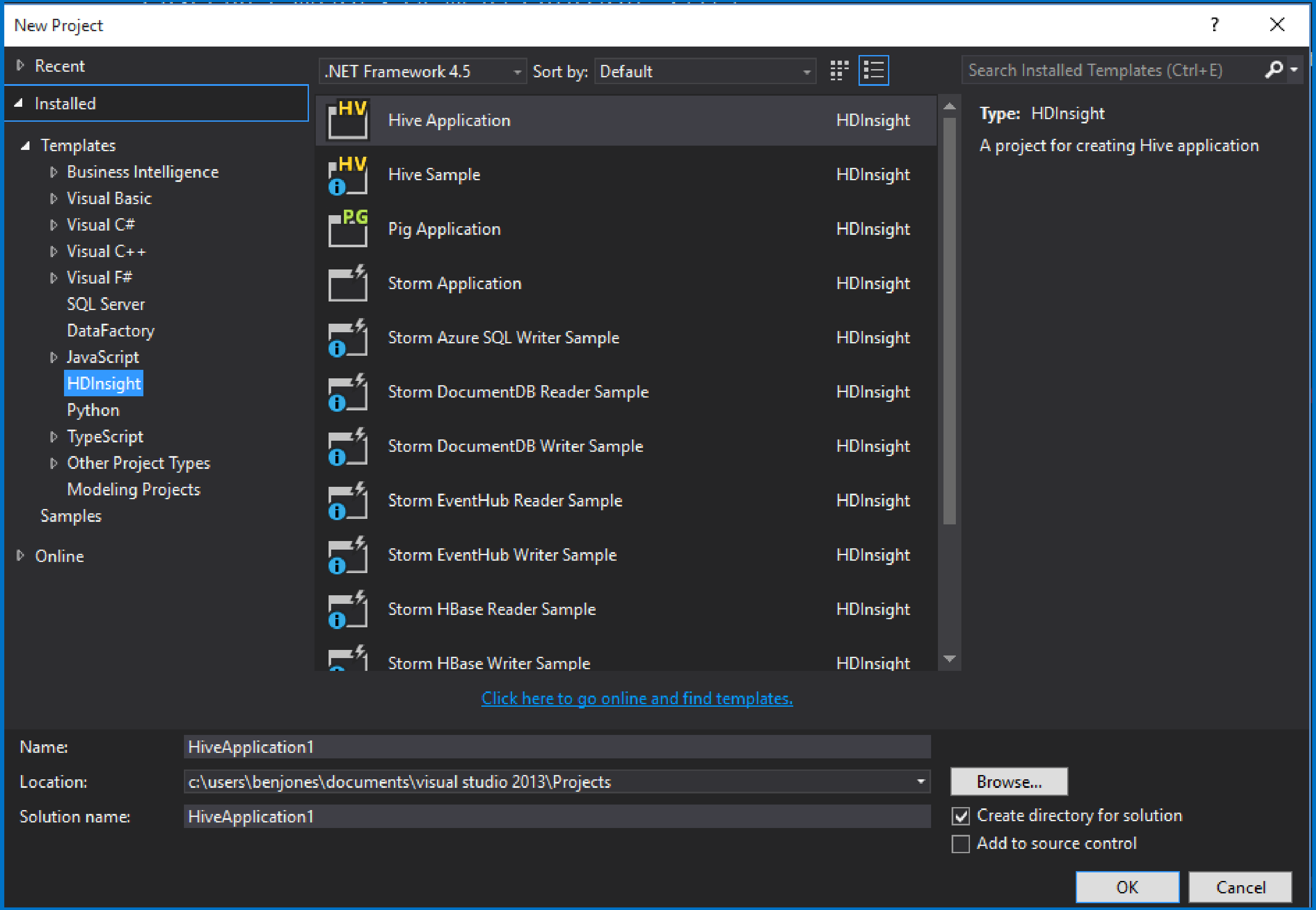
Spout consumes data and emits streams e.g. Twitter API or queueing system such as Kafka, RabbitMQ



Bolt consumes streams and performs CEP / read / writes to output

Streams contain tuples (named list of values) e.g. {Rich, Franz, Kate}

Design & Deployment



New Project

Hive Application	HDInsight
Hive Sample	HDInsight
Pig Application	HDInsight
Storm Application	HDInsight
Storm Azure SQL Writer Sample	HDInsight
Storm DocumentDB Reader Sample	HDInsight
Storm DocumentDB Writer Sample	HDInsight
Storm EventHub Reader Sample	HDInsight
Storm EventHub Writer Sample	HDInsight
Storm HBase Reader Sample	HDInsight
Storm HBase Writer Sample	HDInsight

Type: HDInsight

A project for creating Hive application

Click here to go online and find templates.

Name: HiveApplication1

Location: c:\users\benjones\documents\visual studio 2013\Projects

Solution name: HiveApplication1

☑ Create directory for solution

□ Add to source control

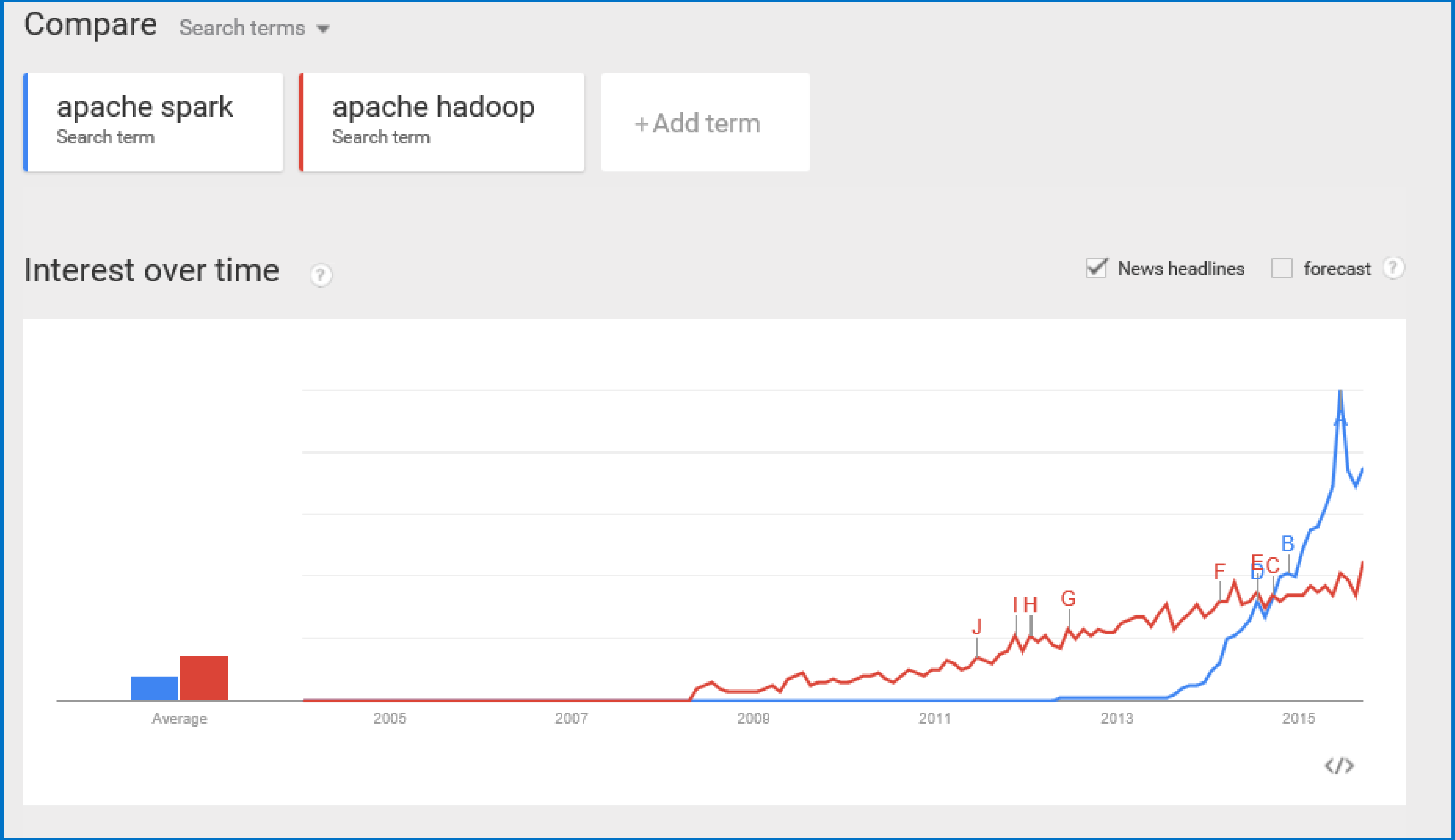
Native Project Support

+

APIs

The Rise of Spark

Spark and



Compare Search terms

apache spark
Search term

apache hadoop
Search term

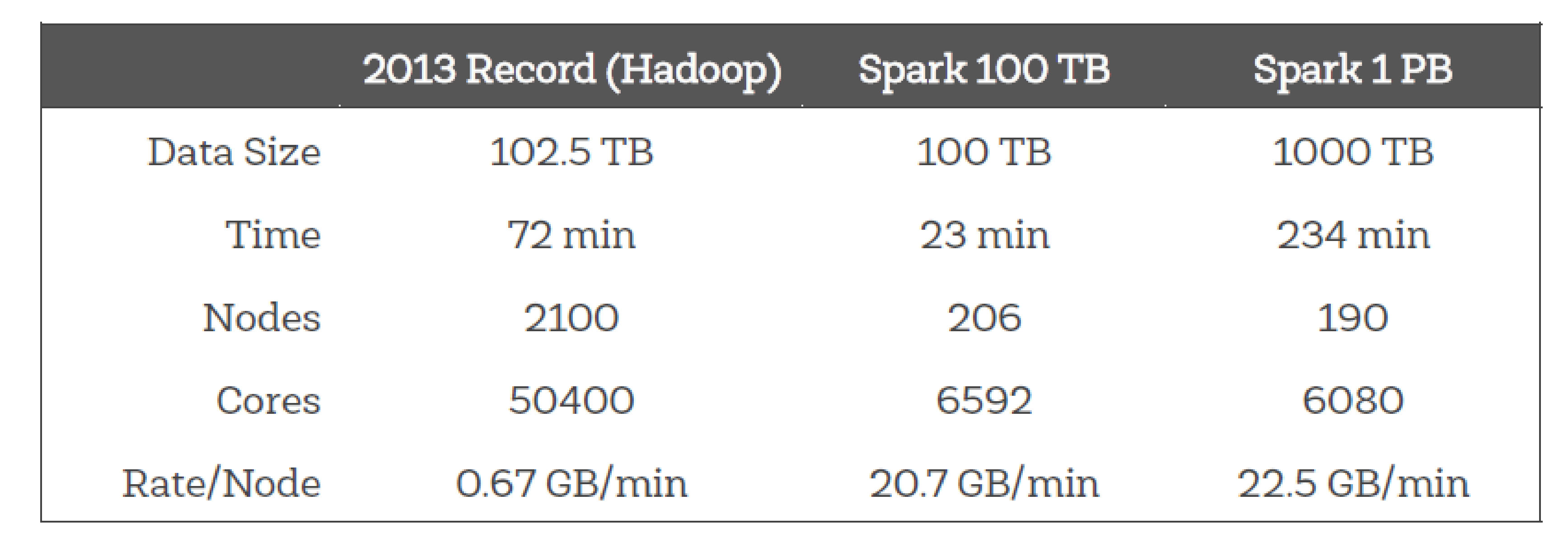
+ Add term

Interest over time

☑ News headlines □ forecast

Spark is fast

Spark is the current (2014) Sort Benchmark winner.
3x faster than 2013 winner (Hadoop).



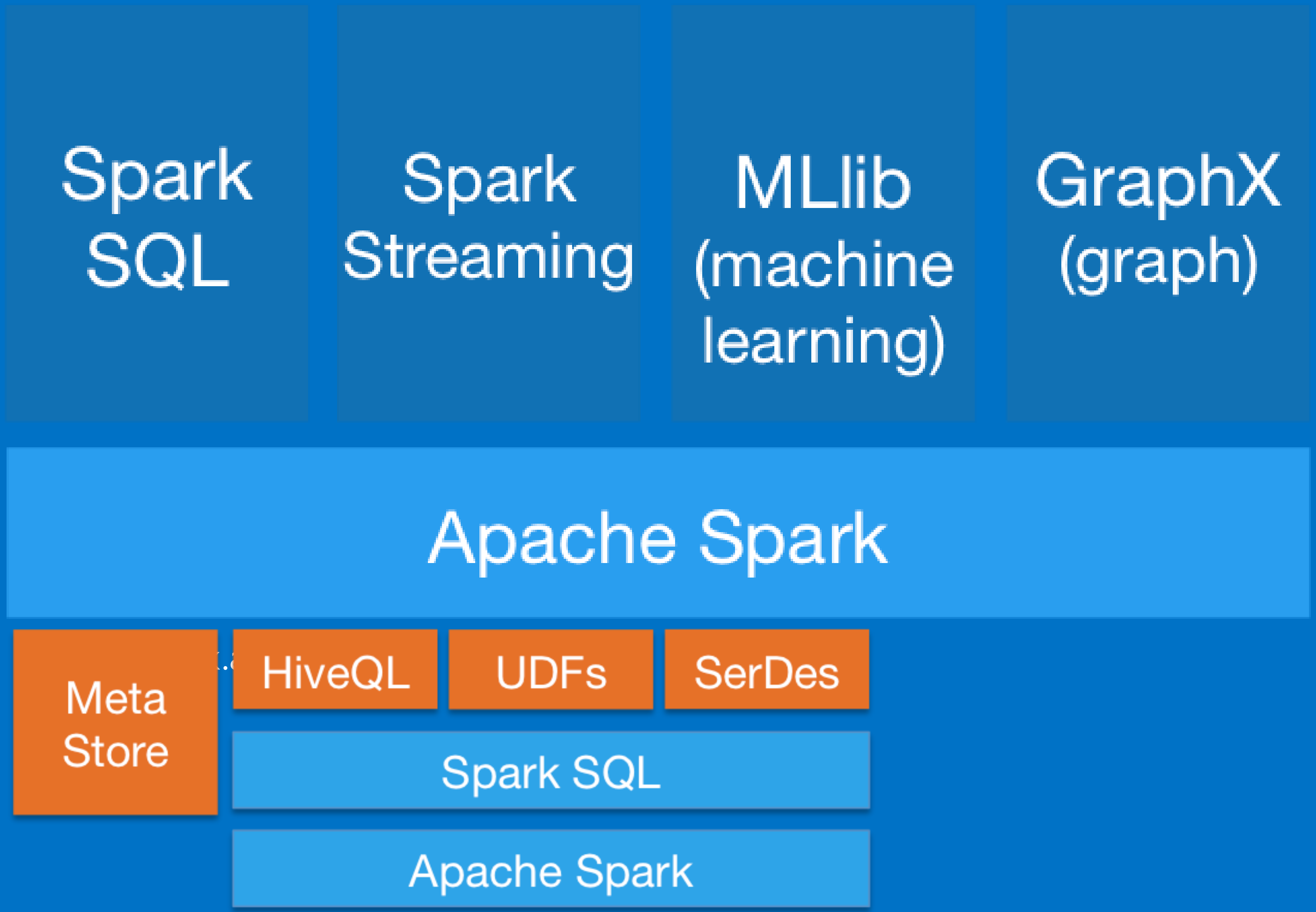
	2013 Record (Hadoop)	Spark 100 TB	Spark 1 PB
Data Size	102.5 TB	100 TB	1000 TB
Time	72 min	23 min	234 min
Nodes	2100	206	190
Cores	50400	6592	6080
Rate/Node	0.67 GB/min	20.7 GB/min	22.5 GB/min

"Spark officially sets a new record in large-scale sorting"

https://databricks.com/blog/2014/11/05/spark-officially-sets-a-new-record-in-large-scale-sorting.html

What is Spark ?

- Spark Unifies:
  - ☆ Batch Processing
  - ☆ Real-time processing
  - ☆ Stream Analytics
  - ☆ Machine Learning
  - ☆ Interactive SQL
  - ☆ Power BI!



Spark on Azure HDInsight

Differentiators

Enterprise-Ready

- Spark as a fully managed Service
- Enterprise Support
- Ease of provisioning

Streaming Capabilities

- Azure integration
- First Class Connector for Azure Event Hubs

Data Insight

- ML libraries & interactive experience through the Notebooks
- Native Integration and Exploration with Power BI + others

Flexibility and Choice

- Node Sizing

High Performance Storage

- SSD Caching

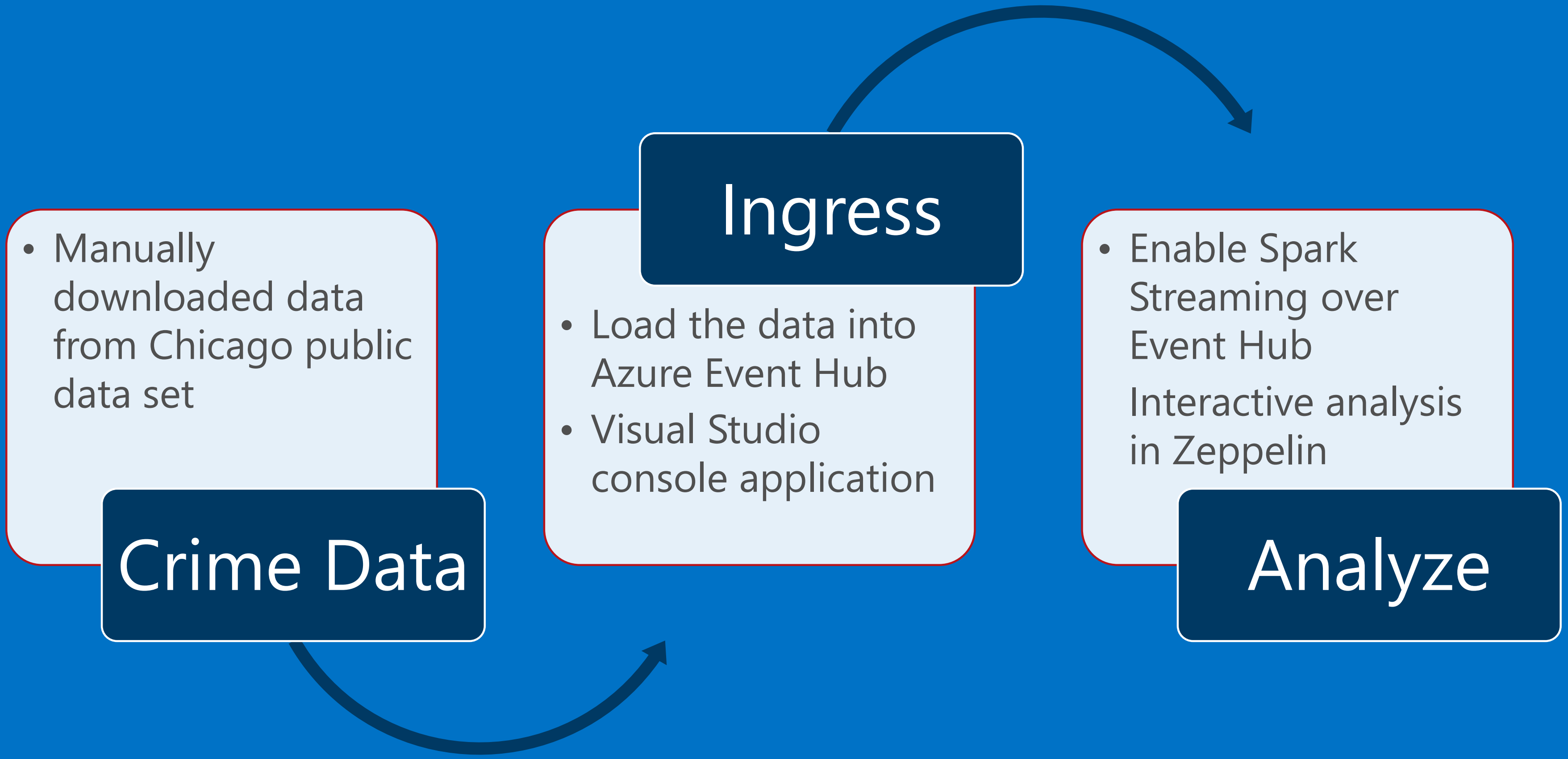
Spark Streaming vs Storm

Spark Streaming differs in a number of ways:

1. Workload - Spark Streaming implements a method for "batching" incoming updates vs. individual events (Storm)
2. Latency - seconds (Spark) vs. sub-second (Storm)
3. Fault Tolerance - exactly once vs at least once

Zaharia, Matei, et al. "Discretized streams: Fault-tolerant streaming computation at scale." Proceedings of the Twenty-Fourth ACM Symposium on Operating Systems Principles. ACM, 2013.

Spark Streaming Demo



- Manually downloaded data from Chicago public data set

Crime Data

Ingress

- Load the data into Azure Event Hub
- Visual Studio console application

- Enable Spark Streaming over Event Hub

  Interactive analysis in Zeppelin

Analyze

Chicago Crime data (2001 to Present) https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Spark Cluster Sizing

Head Node

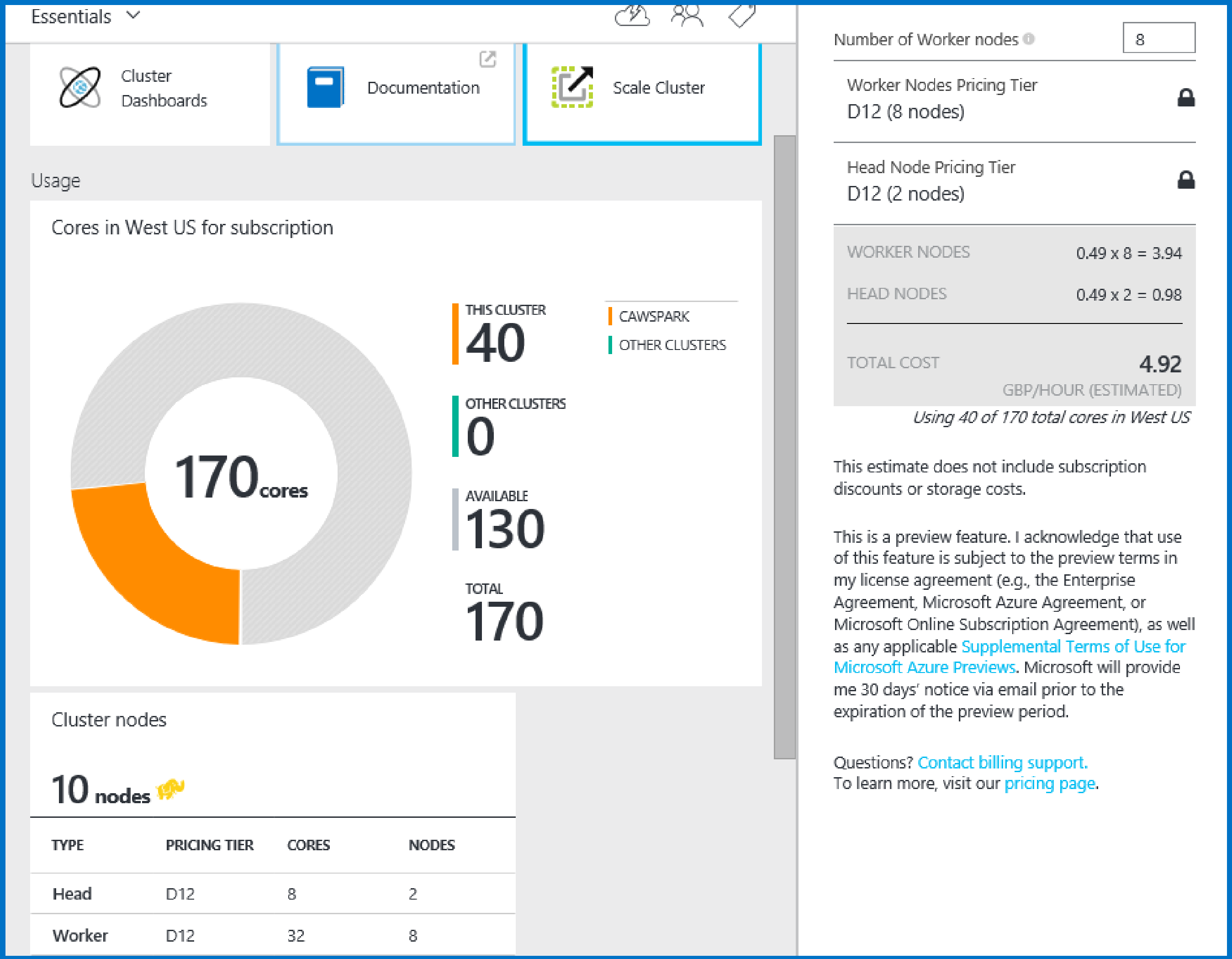
D12 x 2

4 core, 28GB memory, 200 GB SSD

Worker Nodes

D12 x 8

4 core, 28GB memory, 200 GB SSD



Usage

Cores in West US for subscription

THIS CLUSTER 40

OTHER CLUSTERS 0

AVAILABLE 130

TOTAL 170

Cluster nodes

10 nodes

TYPE	PRICING TIER	CORES	NODES
Head	D12	8	2
Worker	D12	32	8

Number of Worker nodes: 8

Worker Nodes Pricing Tier
D12 (8 nodes)

Head Node Pricing Tier
D12 (2 nodes)

WORKER NODES	0.49 x 8 = 3.94
HEAD NODES	0.49 x 2 = 0.98
TOTAL COST	4.92 GBP/HOUR (ESTIMATED)

Using 40 of 170 total cores in West US

This estimate does not include subscription discounts or storage costs.

This is a preview feature. I acknowledge that use of this feature is subject to the preview terms in my license agreement (e.g., the Enterprise Agreement, Microsoft Azure Agreement, or Microsoft Online Subscription Agreement), as well as any applicable Supplemental Terms of Use for Microsoft Azure Previews. Microsoft will provide me 30 days' notice via email prior to the expiration of the preview period.

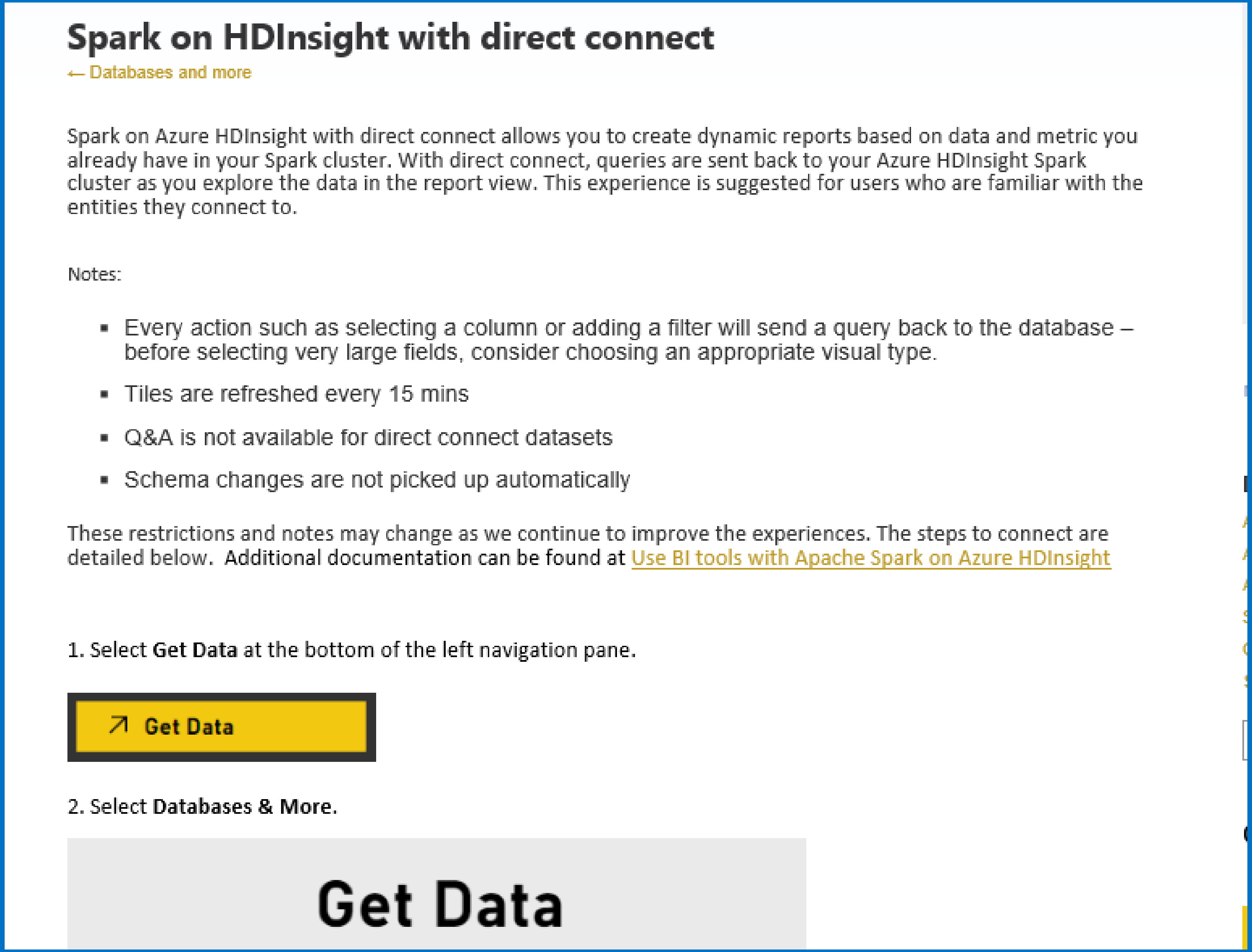
Questions? Contact billing support.
To learn more, visit our pricing page.

http://azure.microsoft.com/en-us/pricing/details/hdinsight/

Demo

Spark Streaming on Azure HDInsight

Spark and Power BI



Spark on HDInsight with direct connect

← Databases and more

Spark on Azure HDInsight with direct connect allows you to create dynamic reports based on data and metric you already have in your Spark cluster. With direct connect, queries are sent back to your Azure HDInsight Spark cluster as you explore the data in the report view. This experience is suggested for users who are familiar with the entities they connect to.

Notes:

- Every action such as selecting a column or adding a filter will send a query back to the database – before selecting very large fields, consider choosing an appropriate visual type.
- Tiles are refreshed every 15 mins
- Q&A is not available for direct connect datasets
- Schema changes are not picked up automatically

These restrictions and notes may change as we continue to improve the experiences. The steps to connect are detailed below. Additional documentation can be found at Use BI tools with Apache Spark on Azure HDInsight

1. Select **Get Data** at the bottom of the left navigation pane.

2. Select **Databases & More.**

https://support.powerbi.com/knowledgebase/articles/654094

Decision Points

	Stream Analytics	HDInsight Storm	HDInsight Spark
Multi-Tenant Service	Yes	No	No
Deployment Model	PaaS	PaaS*	PaaS*
Extensibility	Low	High	High
Deployment Complexity	Low	Low*	Low*
Cost	Low	Med	Med
Open Source Support	No	Yes	Yes
Programmability	SQL*	.NET, Java, Python	SparkSQL, Scala, Python, Java…
Power BI Integration	Yes, Native	Rest API	Yes, Native

* Considerations

Want to learn more about Stream Analytics?

Tutorial: Unlocking Real-Time Insights for Your IoT Data
1:00 – 3:00 PM in the **Rainier Room**

Build out an end-to-end stream processing solution over vehicle telemetry

Resources

Overview: Azure Stream Analytics
https://azure.microsoft.com/en-us/documentation/articles/stream-analytics-introduction/

Overview: Storm on HDInsight
https://azure.microsoft.com/en-us/documentation/articles/hdinsight-storm-overview

Comparison of Apache Storm and Azure Stream Analytics
https://azure.microsoft.com/en-us/documentation/articles/stream-analytics-comparison-storm/

Overview: Apache Spark on Azure HDInsight
https://azure.microsoft.com/en-us/documentation/articles/hdinsight-apache-spark-overview/

Spark Streaming: Process events from Azure Event Hubs with Apache Spark on HDInsight
https://azure.microsoft.com/en-us/documentation/articles/hdinsight-apache-spark-csharp-apache-zeppelin-eventhub-streaming

Build Machine Learning applications using Apache Spark on Azure HDInsight
https://azure.microsoft.com/en-us/documentation/articles/hdinsight-apache-spark-ipython-notebook-machine-learning/

Microsoft

Benjamin Wright-Jones

benjones@microsoft.com

Simon Lidberg

simonlid@microsoft.com