Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

ROCK THE CODE TOUR 2017 | CHICAGO

Big Data on Azure

David Giard
Microsoft Technical Evangelist
dgiard@Microsoft.com
Davidgiard.com
@davidgiard

David Giard
Microsoft Technical Evangelist
dgiard@Microsoft.com
Davidgiard.com
@davidgiard

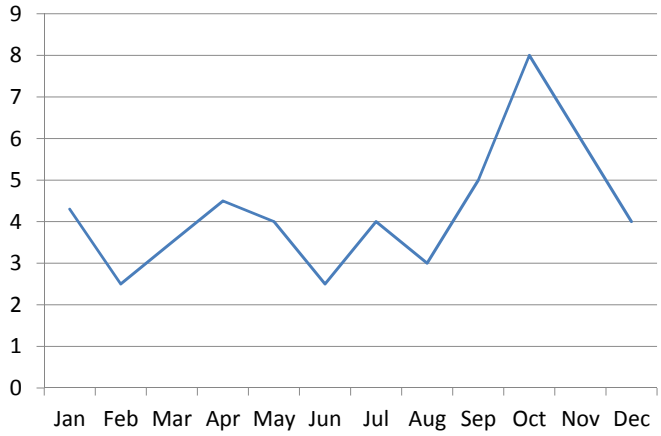
Cloud Computing

Host some or all
of your data or application
on a third-party server
in a highly-scalable,
highly-reliable way

Advantages of Cloud Computing

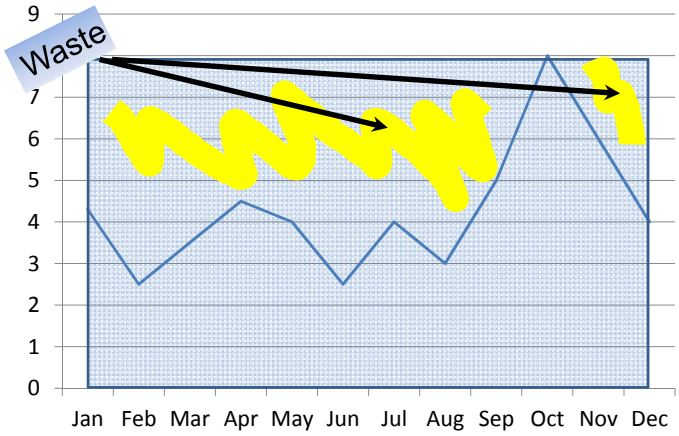
- Lower capital costs
- Flexible operating cost (Rent vs Buy)
- Platform as a Service
- Freedom from infrastructure / hardware
- Redundancy
- Automatic monitoring and failover

Demand and Capacity



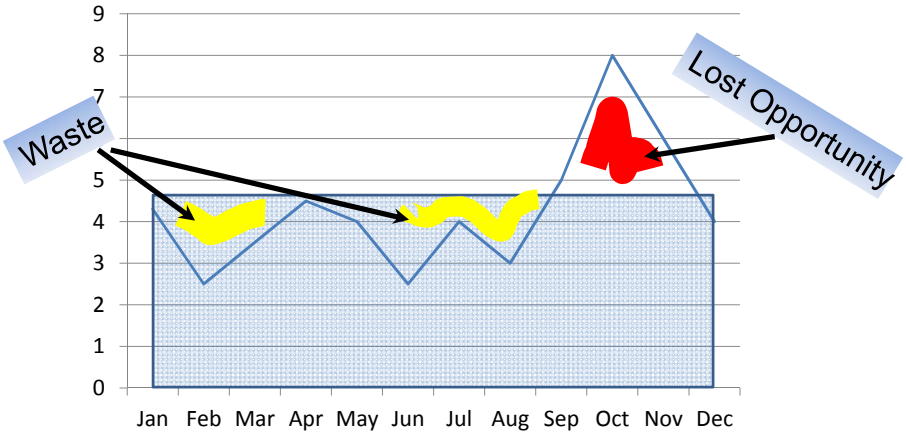
Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

Demand and Capacity



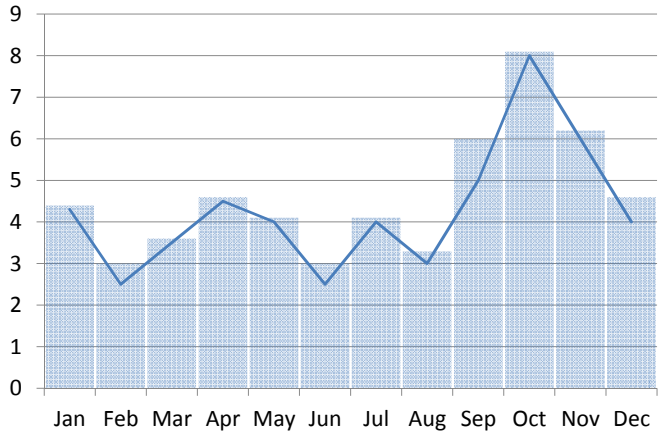
Visual Studio LIVE!

Demand and Capacity



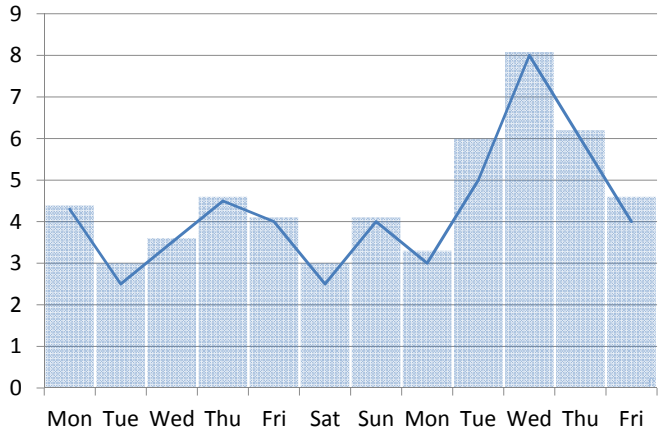
Visual Studio LIVE!

Demand and Capacity



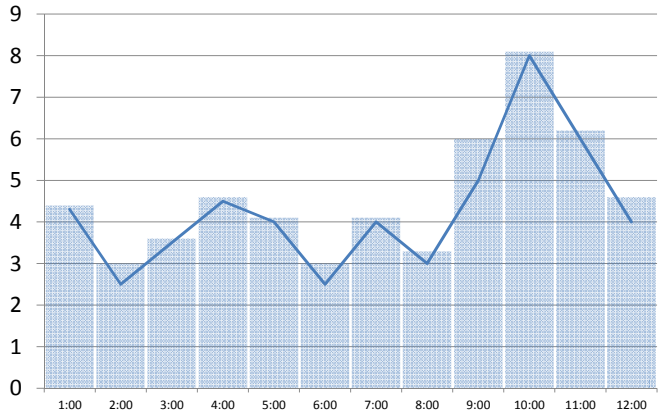
Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

Demand and Capacity



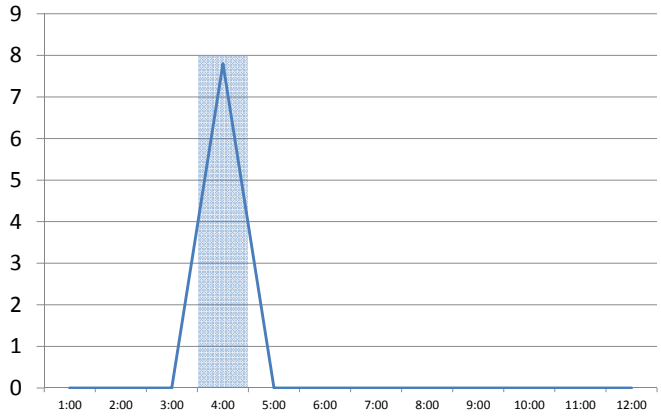
Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

Demand and Capacity

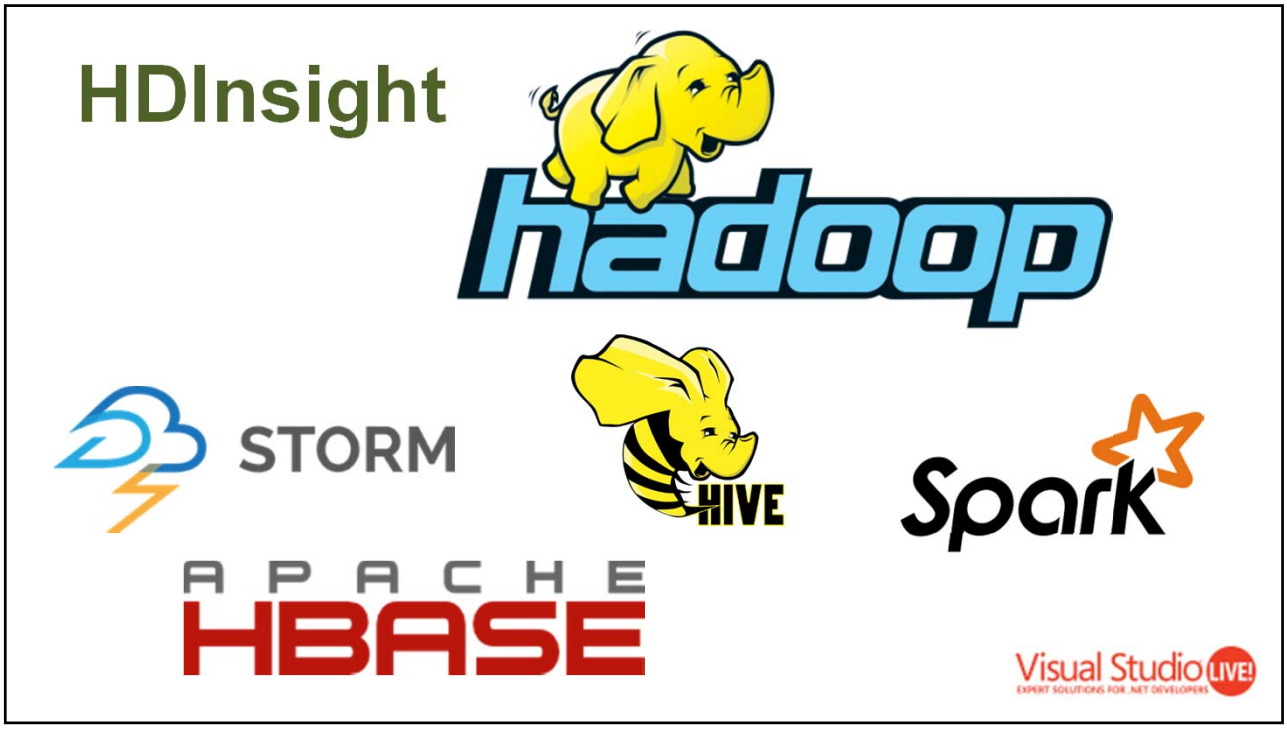


Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

Big Data Demand



Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

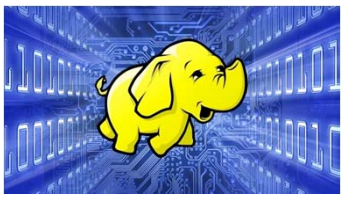


Azure HDInsight

- Microsoft Azure's big-data solution using Hadoop
  - Open-source framework for storing and analyzing massive amounts of data on clusters built from commodity hardware
  - Uses Hadoop Distributed File System (HDFS) for storage
- Employs the open-source Hortonworks Data Platform implementation of Hadoop
  - Includes HBase, Hive, Pig, Storm, Spark, and more
- Integrates with popular BI tools
  - Includes Power BI, Excel, SSAS, SSRS, Tableau

Apache Hadoop on Azure

- Automatic cluster provisioning and configuration
  - Bypass an otherwise manual-intensive process
- Cluster scaling
  - Change number of nodes without deleting/re-creating the cluster
- High availability/reliability
  - Managed solution - 99.9% SLA
  - HDInsight includes a secondary head node
- Reliable and economical storage
  - HDFS mapped over Azure Blob Storage
  - Accessed through “wasb://” protocol prefix

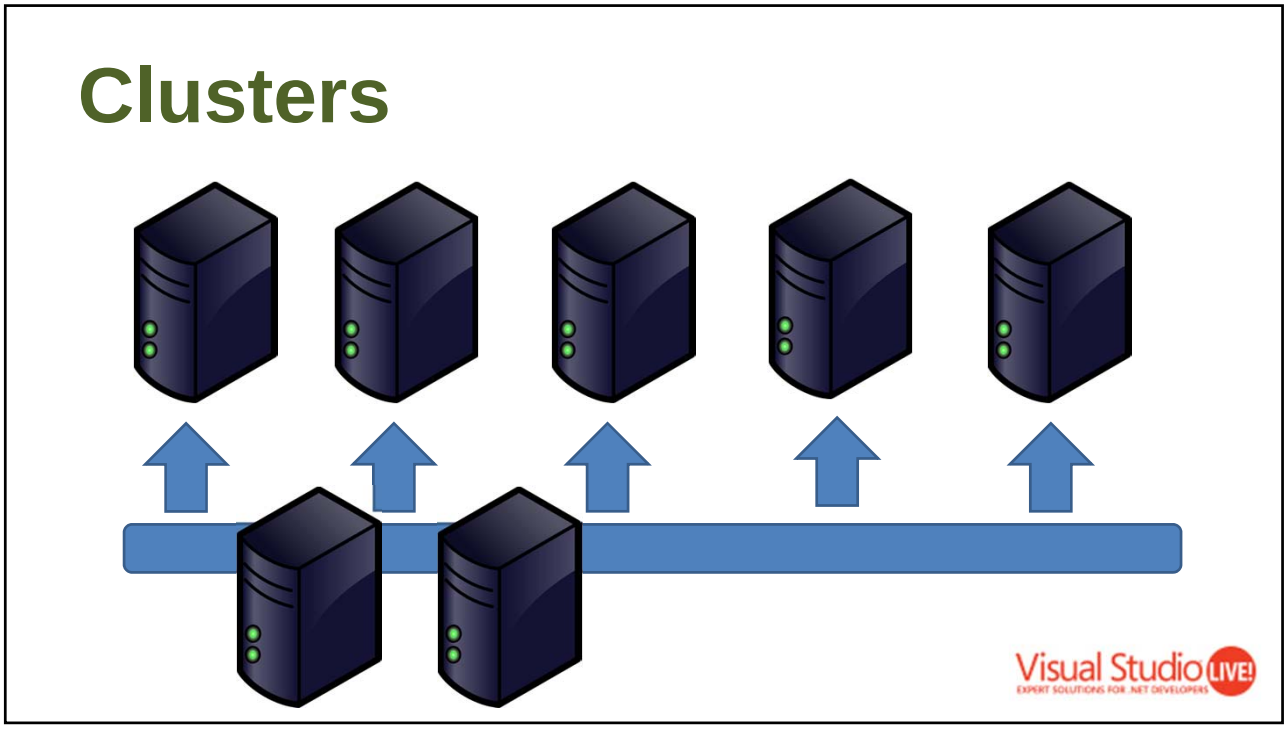


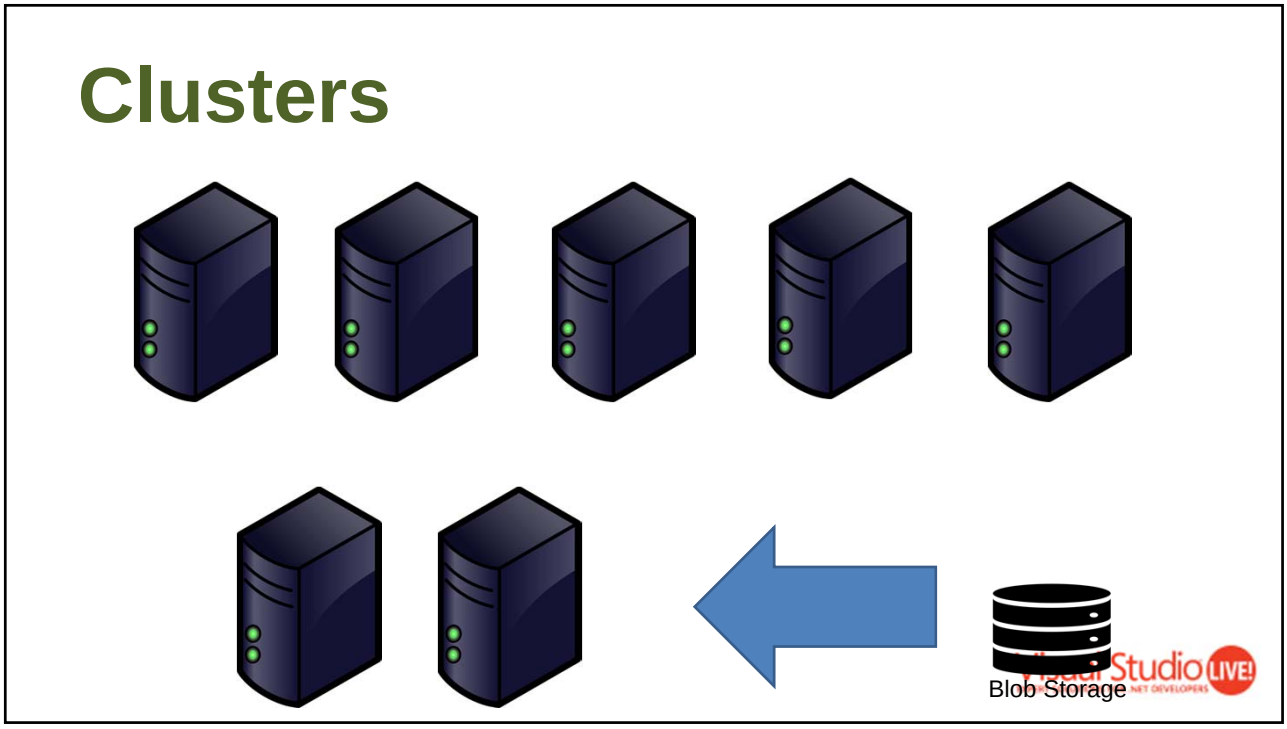
Visual Studio LIVE!

Lambda Architecture

- Batch Layer
- Speed Layer
- Serving Layer

Visual Studio LIVE!

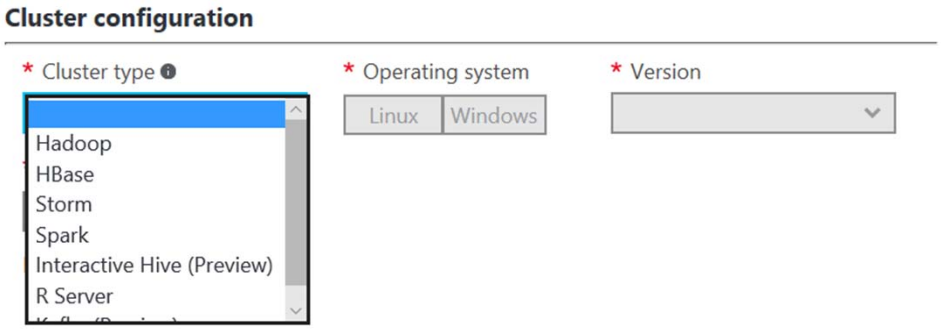

Clusters
Visual Studio LIVE!


Clusters
Blob Storage
Visual Studio LIVE!

HDInsight Cluster Types

- Hadoop: Query workloads
  - Reliable data storage, simple MapReduce
- HBase: NoSQL workloads
  - Distributed database offering random access to large amounts of data
- Apache Storm: Stream workloads
  - Real-time analysis of moving data streams
- Apache Spark: High-performance workloads
  - In-memory parallel processing

Cluster Creation



Cluster Creation

```
{
  "$schema": "https://schema.management.azure.com/schemas/2015-01-01/deploymentTemplate.json#",
  "contentVersion": "1.0.0.0",
  "parameters": {
    "clusterName": {
      "type": "string",
      "metadata": {
        "description": "The name of the HDInsight cluster to create."
      }
    },
    "clusterLoginUserName": {
      "type": "string",
      "defaultValue": "admin",
      "metadata": {
        "description": "These credentials can be used to submit jobs to the cluster and to log into cluster
dashboards."
      }
    },
    "clusterLoginPassword": {
```

Demo

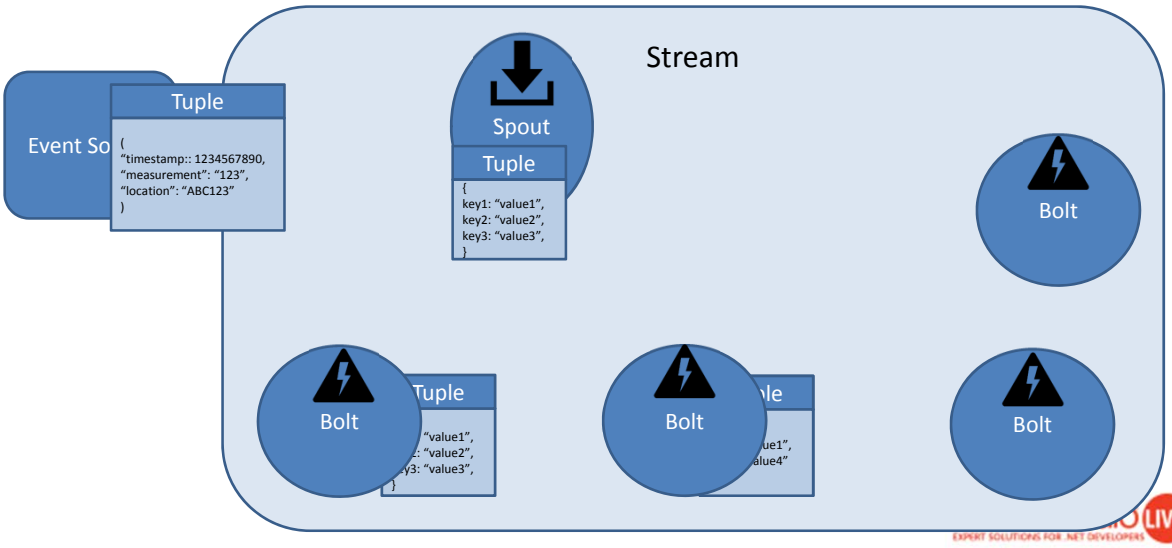
Storm

- Apache Storm is a distributed, fault-tolerant, open-source computation system that allows you to process data in real-time with Hadoop.
- Apache Storm on HDInsight allows you to create distributed, real-time analytics solutions in the Azure environment by using Apache Hadoop.
- Storm solutions can also provide guaranteed processing of data, with the ability to replay data that was not successfully processed the first time.
- Ability to write Storm components in **C#, JAVA and Python**.
- Azure Scale up or Scale down without an impact for running Storm topologies.
- Ease of provision and use in Azure portal.
- Visual Studio project templates for Storm apps

Storm

- Apache Storm apps are submitted as Topologies.
- A topology is a graph of computation that processes streams
- Stream: An unbound collection of tuples. Streams are produced by spouts and bolts, and they are consumed by bolts.
- Tuple: A named list of dynamically typed values.
- Spout: Consumes data from a data source and emits one or more streams.
- Bolt: Consumes streams, performs processing on tuples, and may emit streams. Bolts are also responsible for writing data to external storage, such as a queue, HDInsight, HBase, a blob, or other data store.
- Nimbus: JobTracker in Hadoop that distribute jobs, monitoring failures.

Apache Storm Topology



Demo

Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

APACHE HBASE

Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS
@DavidGiard

HBase

- Apache HBase is an open-source, **NoSQL** database that is built on Hadoop and modeled after Google BigTable.
- HBase provides random access and strong consistency for large amounts of **unstructured** and **semistructured** data in a **schemaless** database organized by **column families**
- Data is stored in the rows of a table, and data within a row is grouped by column family.
- The open-source code scales **linearly** to handle **petabytes** of data on thousands of nodes. It can rely on data redundancy, batch processing, and other features that are provided by distributed applications in the Hadoop ecosystem.

HBase

- HBase Commands:
  - create ➔ Equivalent to create table in T-SQL
  - get ➔ Equivalent to select statements in T-SQL
  - put ➔ Equivalent to update, Insert statement in T-SQL
  - scan ➔ Equivalent to select (no where condition) in T-SQL
  - delete ➔ Equivalent to delete in T-SQL
- **HBase shell** is your query tool to execute in CRUD commands to a HBase cluster.
- Data can also be managed using the **HBase C# API**, which provides a client library on top of the **HBase REST API**.
- An HBase database can also be queried by using **Hive**.

HBase

Column family “a”

Column family “b”

Column family “c”

RowKey	a:1	a:2	a:3	a:4	b:1	b:2	c:numA
982069	10	20	30	40	5	7	4
926025	9	11	21		4	9	3
254114	11	15	22	35	7	11	4
881514	8	14			2	3	2

HBase

Column family “temperature”

Column family “pressure”

Column family “avg_temp”

RowKey	a:1	a:2	a:3	a:4	b:1	b:2	c:numA
982069	10	20	30	40	5	7	4
926025	9	11	21		4	9	3
254114	11	15	22	35	7	11	4
881514	8	14			2	3	2

Demo

Hive

- Apache Hive is a **data warehouse system for Hadoop**, which enables data summarization, querying, and analysis of data by using **HiveQL** (a query language similar to SQL).
- Hive understands how to work with structured and semi-structured data, such as text files where the fields are delimited by specific characters.
- Hive also supports custom **serializer/deserializers** for complex or irregularly structured data.
- Hive can also be extended through **user-defined functions (UDF)**.
- A UDF allows you to implement functionality or logic that isn't easily modeled in HiveQL.

HiveQL

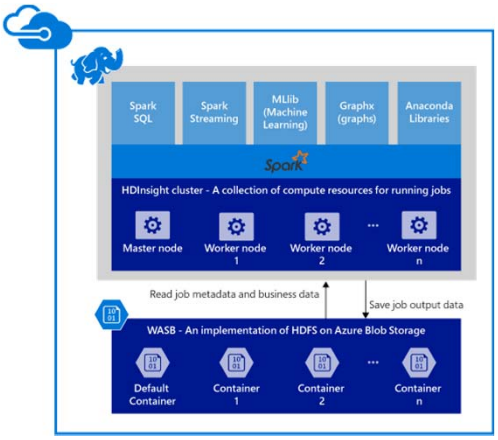
```
# Number of Records
SELECT COUNT(1) FROM www_access;

# Number of Unique IPs
SELECT COUNT(1) FROM ( \
  SELECT DISTINCT ip FROM www_access \
) t;

# Number of Unique IPs that Accessed the Top Page
SELECT COUNT(distinct ip) FROM www_access \
  WHERE url='/';

# Number of Accesses per Unique IP
SELECT ip, COUNT(1) FROM www_access \
  GROUP BY ip LIMIT 30;

# Unique IPs Sorted by Number of Accesses
SELECT ip, COUNT(1) AS cnt FROM www_access \
  GROUP BY ip
  ORDER BY cnt DESC LIMIT 30;

# Number of Accesses After a Certain Time
SELECT COUNT(1) FROM www_access \
  WHERE TD_TIME_RANGE(time, "2011-08-19", NULL, "PDT")
```

Apache Spark

- Interactive manipulation and visualization of data
  - Scala, Python, and R Interactive Shells
  - Jupyter Notebook with PySpark (Python) and Spark (Scala) kernels provide in-browser interaction
- Unified platform for processing multiple workloads
  - Real-time processing, Machine Learning, Stream Analytics, Interactive Querying, Graphing
- Leverages in-memory processing for really big data
  - Resilient distributed datasets (RDDs)
  - APIs for processing large datasets
  - Up to 100x faster than MapReduce

Spark Components on HDInsight

- Spark Core
  - Includes Spark SQL, Spark Streaming, GraphX, and MLlib
- Anaconda
- Livy
- Jupyter Notebooks
- ODBC Driver for connecting from BI tools (Power BI, Tableau)



Jupyter Notebooks on HDInsight

- Browser-based interface for working with text, code, equations, plots, graphics, and interactive controls in a single document.
- 

Demo

Visual Studio LIVE!
EXPERT SOLUTIONS FOR .NET DEVELOPERS

Items of Note About HDInsight

- There is no “suspend” on HDInsight clusters
  - Provision the cluster, do work, then delete the cluster to avoid unnecessary charges
  - Storage can be decoupled from the cluster and reused across deployments
- Can deploy from the portal, but often scripted in practice
  - Easier/repeatable creation and deletion

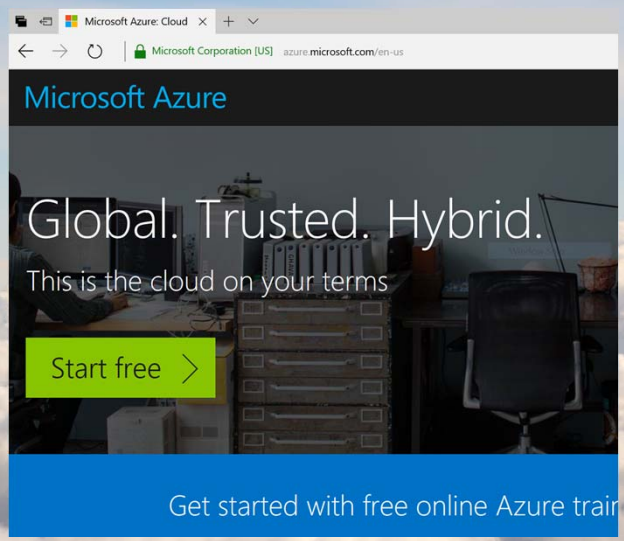
Links

www.slideshare.net/dgiard/big-data-on-azure-70554456

github.com/MSFTImagine/computerscience/tree/master/Workshop/7.%20HDInsight

Get Started

azure.com




Thank you!
@DavidGiard