CapTech

HDInsight > Hadoop

October 12, 2017

Others Talk,
We Listen.

2 Introduction

- Mark Hudson
  - >20 years mixing technology with data
  - >10 years with CapTech
  - Microsoft Certified IT Professional – Business Intelligence
  - Member of the Richmond SQL Server User Group
  - Email: mhudson@captechconsulting.com
  - Twitter: @HMarkHudson

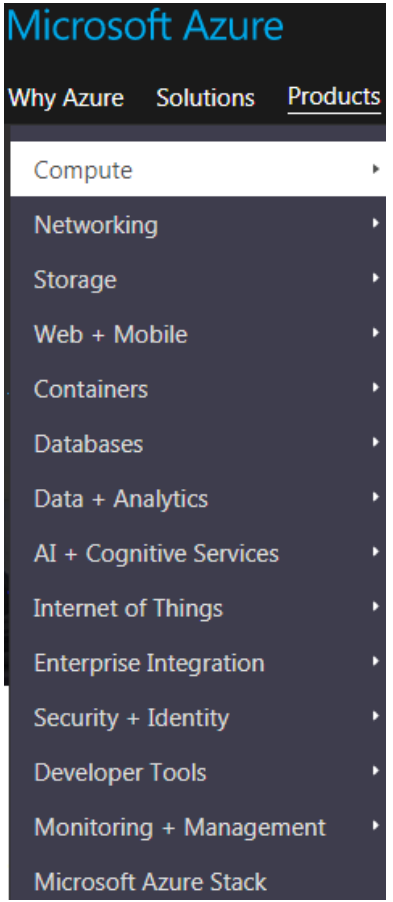
- CapTech
  - Headquarters – Richmond
  - Offices – Washington, Charlotte, Philadelphia, Baltimore, Chicago, Orlando, Atlanta, Denver
  - Local, national, and international clients
  - Microsoft Gold Partner
  - Web: www.captechconsulting.com
  - Twitter: @CapTechListens

3 Agenda

- Azure, Storage, and Data + Analytics
- HDInsight
  - Hadoop Demo
  - Spark Demo
- Learning Material
- Questions

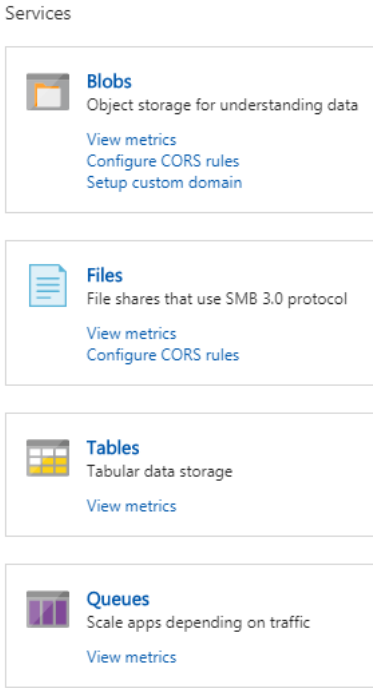
4 Microsoft Azure

- Microsoft's Pay-As-You-Go Cloud Offerings
  - IaaS – Servers, storage, networks, data centers
  - PaaS – IaaS + middleware, tools, services
  - SaaS – IaaS + PaaS + applications
- Cloud Market Growth
  - Gartner estimates 18-30% CAGR spending through 2021
- Cloud Market Share
  - Amazon Web Service – 47%
  - Azure – 10%
  - Google – 4%
  - Others – 39%



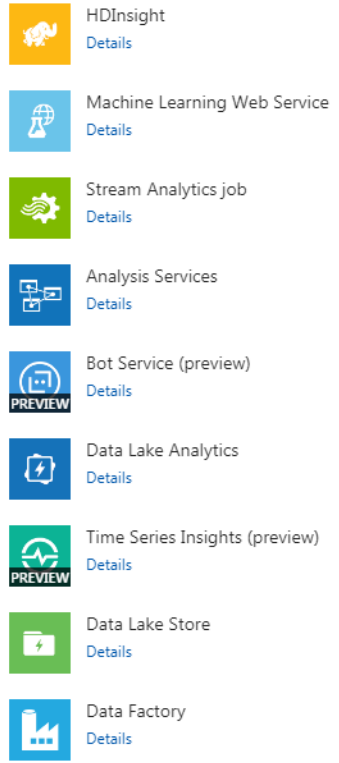
5 Azure Storage

- Types of Azure Storage
  - Blob – Flexible, accessible files (e.g., flat, formatted, image)
    - Tools – AZCopy, Azure Storage Explorer, Cloudberry Explorer, Azure Command Line
  - File – File share from anywhere using Server Message Block (SMB) and Shared Access Signature (SAS) token
  - Table – NoSQL key-value pair (i.e., CosmosDB)
  - Queue – Store and retrieve up to 64KB messages
  - Disk – Optimized for virtual machines
  - More options – Standard/Premium, Cool/Hot, Local/Zone/Geo redundant
- Data Lake Store
  - Separate data tool from Blob Storage and Data Lake Analytics
  - Optimized storage for big data analytics
  - Distinct management options



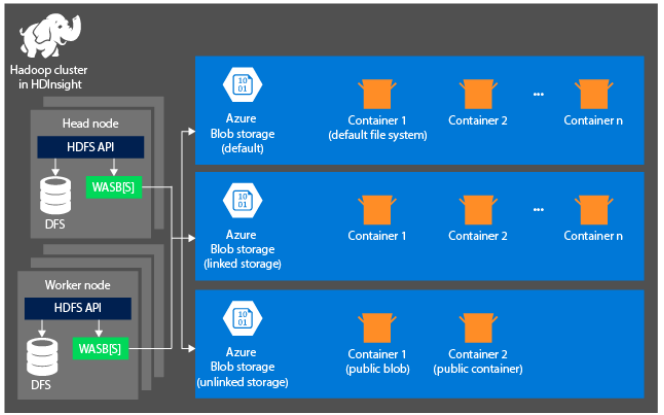
6 Azure Data + Analytics

- Generally Available Services
  - HDInsight – Fully-managed cloud Apache Hadoop and more
  - Machine Learning Web Service – Deployed AzureML predictive models
  - Stream Analytics Job – Real-time analytics on streaming data
  - Cognitive Services – Publicly available intelligence APIs for subscription
  - Data Lake Analytics – Massively parallel data transformation and processing programs
  - Data Lake Store – Massively scalable, HDFS standard for massively-parallel analytics
  - Data Factory – Data integration service automating data movement and transformation
  - Data Catalog – Enterprise data asset metadata
- Preview Services
  - Bot Service – Purpose-built for bot development … fun
  - Time Series Insights – Explore and analyze IoT events



7 Azure HDInsights

- Generally Available Cluster Types
  - Hadoop – Reliable, scalable, distributed computing
  - HBase – NoSQL database provides random access to large amounts of data
  - Storm – Real-time computation system for processing large streams of data
  - Spark – Parallel in-memory processing supports performance of big-data analysis
  - R Server – Parallel, distributed R
- Preview Cluster Types
  - Kafka – Streams data pipelines with message-queue functionality to publish and subscribe to data streams
  - Interactive Hive – In-memory caching for faster queries
- Separate storage from computing
  - Facilitates data reuse
  - Reduces cost of storing data
  - Computing instances “as-needed”



8 HDInsight Hadoop

- Apache Hadoop (Hortonworks Data Platform)
  - File system (HDFS) enabling sequential data access
- Provides different cluster versions containing different component versions
  - Hortonworks Data Platform, Hadoop and Yarn, Tez, Pig, Hive and HCatalog, Hive2, Tez Hive2

Component	HDInsight 3.6 (default)	HDInsight 3.5	HDInsight 3.4	HDInsight 3.3	HDInsight 3.2	HDInsight 3.1
Hortonworks Data Platform	2.6	2.5	2.4	2.3	2.2	2.1.7
Apache Hadoop and YARN	2.7.3	2.7.3	2.7.1	2.7.1	2.6.0	2.4.0
Apache Tez	0.7.0	0.7.0	0.7.0	0.7.0	0.5.2	0.4.0
Apache Pig	0.16.0	0.16.0	0.15.0	0.15.0	0.14.0	0.12.1
Apache Hive and HCatalog	1.2.1	1.2.1	1.2.1	1.2.1	0.14.0	0.13.1
Apache Hive2	2.1.0	-	-	-	-	-
Apache Tez Hive2	0.8.4	-	-	-	-	-

9 HDInsight Hadoop Demo

- Azure HDInsight
  - https://portal.azure.com/#create/Microsoft.HDInsightCluster
- Visual Studio 2017
- Lahman's Baseball Database
  - http://www.seanlahman.com/baseball-archive/statistics/

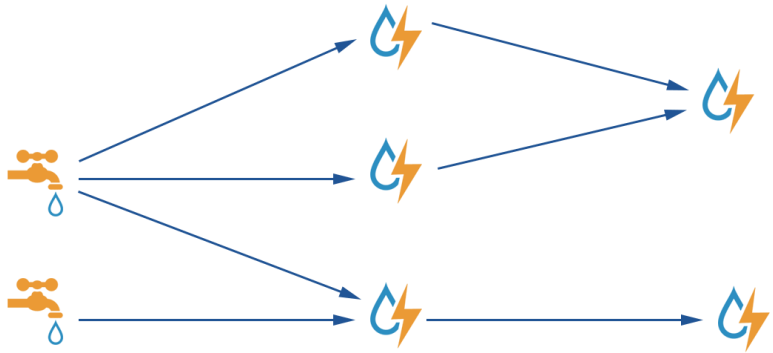
Please delete your demo.

10 HDInsight HBase

- Apache HBase
  - NoSQL database enabling random read/write data access
- Stores data in Blob Storage or Data Lake Store
  - Runs on Hadoop HDFS
- Structured as tables of rows in column families
  - Key Value data store
- High-performance of targeted query results
  - Depends on data keyed on same column(s)

11 HDInsight Storm

- Apache Storm
- Processes streams of data in real time
  - Spouts ingest data from source
  - Bolts process and write data
- Integrates with other Azure services
  - Data Lake Store
  - Storage
  - Event Hubs
  - SQL Database
  - Cosmos DB
  - HBase



HDInsight Spark

- Apache Spark
  - Framework supports in-memory processing to boost performance of big-data analytics
  - Compatible with Azure Storage Blob (WASB) as well as Azure Data Lake Store
  - Supports Scala and PySpark programming languages

13 HDInsight Spark Demo

- Azure HDInsight
  - https://portal.azure.com/#create/Microsoft.HDInsightCluster
- PySpark
- Lahman's Baseball Database
  - http://www.seanlahman.com/baseball-archive/statistics/

Please delete your demo.

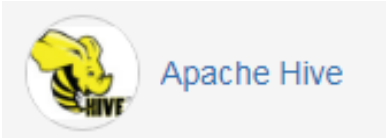
14 HDInsight R Server

- Open-source R capabilities
  - 8000+ R packages available
- Persist data or code from Blob Storage, File System Storage, or Data Lake Store
- Code using SSH edge node login with R, RStudio installed on edge node, or R Tools for Visual Studio

15 HDInsight Kafka - Preview

- Apache Kafka
- Streams data records in topics

16 HDInsight Interactive Hive - Preview



- Apache Hive
- Low Latency Analytical Processing (LLAP)
  - AKA Live Long and Process
- Leverages in-memory caching, pre-fetching, daemon
  - Improved performance over Hadoop cluster
- Subset of Hadoop Functionality
  - No MapReduce, Pig, Sqoop, Oozie, etc.
  - Accessible only through the Ambari Hive view, Beeline, and Hive ODBC

17 Azure HDInsight Learning Material

Learn Hadoop on HDInsight

https://azure.microsoft.com/en-us/documentation/learning-paths/hdinsight-self-guided-hadoop-training/

Microsoft Virtual Academy Courses

https://mva.microsoft.com/training-topics/cloud-app-development#!jobf=Developer&lang=1033

18 Questions