Microsoft

Cortana Analytics Workshop

Sept 10 – 11, 2015 • MSCC

The “Big Data” of the Cortana Analytics Suite

Lance Olson
Partner Group Program Manager

Microsoft

Agenda

Big Data As Part of Cortana Analytics

Business Scenarios For Big Data

What To Do Next

From data to decisions and actions

The Characteristics of Data Is Changing

Let's Talk About Hadoop

Apache Open Source Project

Highly scalable distributed file system (HDFS)

Distributed processing on data nodes

Data Volumes

Data Variety

Data Velocity

Why Cloud + Hadoop?

Data of all Volume
Variety, Velocity

Massive Compute
and Storage

Deployment
expertise

Speed

Scale

Economics

Time to value

Always Up,
Always On

Open and flexible

Cortana Analytics Suite

Big Data and Advanced Analytics

Big Data as part of Cortana Analytics

Azure HDInsight Introduction – Make Big Data Easier

- Microsoft's cloud Hadoop offering
- 100% open source Apache Hadoop
- Built on the latest releases for Hadoop
- Up and running in minutes with no hardware to deploy
- .NET and Java skills and deep integration to Visual Studio
- Utilize familiar BI tools for analysis including Microsoft Excel
- 99.9% Enterprise Service Level Agreement

Hive to do SQL-Like Queries

- HDInsight provides easy-to-use graphical query interface for Hive
- HiveQL is a SQL-like language (subset of SQL)
- Hive structures include well-understood database concepts such as tables, rows, columns, partitions
- Compiled into MapReduce jobs that are executed on Hadoop

HBase to do NoSQL

- Random, fast (realtime) read/write access to your Big Data.
- Host very large tables (billions of rows X millions of columns) on clusters of commodity hardware.
- Runs on top of the Hadoop Distributed File System (HDFS)
- Provides flexibility in that new columns can be added to column families at any time

Storm for Real-Time Streaming

- Distributed real-time computation system
- Reliably consumes millions of real-time events from a scalable event broker (ie. Kafka, Event Hub)
- Performs time-sensitive computation
- Output to persistent stores, dashboards or devices
- Completely customizable with Java + .NET
- Deeply integrated to Visual Studio

Spark for In-Memory Computation

- Single execution model for multiple tasks (SQL queries, Streaming, Machine Learning, and Graph)
- Processing up to 100x faster performance
- Developer friendly (Java, Python, Scala)
- BI tool of choice (Power BI, Tabelau, Qlik, SAP)
- Notebook experience (Jupyter/iPython, Zeppelin)

Spark for Azure HDInsight

In-memory computation engine – Fully managed

Integration to Visual Studio

- Debug Hive jobs through Yarn logs or troubleshoot Storm topologies
- Visualize jobs, clusters, tables, and storage
- Submit Hive queries, Storm topologies (C# or Java spouts/bolts)
- IntelliSense

HDInsight Built for Linux or Windows

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

Linux

Windows Server

Partner Spotlight: Datameer

Hadoop Data Preparation and Analytics for the Analyst

Partner Spotlight: Trifacta

Hadoop Data Preparation for the Data Developer – Discovery, Assessment, Shaping, Enrichment

Partner Spotlight: AtScale

Analysts Use Traditional BI Tools Against HDInsight

Azure Data Lake Introduction – Make Big Data Easier

- HDFS For the Cloud
- Unlimited Storage, Petabyte Files
- Optimized for Massive Throughput
- High frequency, low latency, read immediately
- Managed and secured

Hadoop Distributed File System (HDFS) For The Cloud

Microsoft Azure Data Lake

- Built from the ground-up as native HDFS
- Integrated w/ HDInsight, Hortonworks, Cloudera
- Accessible to all HDFS compliant projects (Spark, Storm, Flume, Sqoop, Kafka, R, etc.)

Unlimited Storage, Petabyte Files

- Unlimited account sizes
- Individual file sizes from GBs to PBs
- Immediate read/write access

Optimized for Massive Throughput

- Built for running large analytic systems that require massive throughput
- Automatically optimize for any throughput
- Optimized for parallel computation over PBs of data

High frequency, low latency, read immediately

- High volumes of small writes at low latency
- Optimized for near real-time scenarios like IoT, real-time fraud detection, clickstream analysis, etc.

Manage and secure your data assets

- Monitor performance, receive alerts, and audit usage
- Azure Active Directory integration for identity and access management over all of your data

Agenda

Big Data As Part of Cortana Analytics

Business Scenarios For Big Data

What To Do Next

Neudesic partnered with one of the nation's largest utility companies that recently deployed Smart Utility Meters for power customers, nearly a million meters sending usage data every 15 minutes.

The result: an Azure hybrid big data processing solution that enabled the customer to perform gap analytics: a process for identifying gaps that exist in the power usage readings, over 7x faster than their previous solution! Billions of Smart Meter reads get processed to identify the nature and duration of the gaps to mitigate revenue losses.

Cortana Analytics Key Verticals

Retail

Industry Overview

- $22 trillion global revenue ($4.7 trillion U.S.)
- 4% annual growth rate
- 3.7 million businesses
- 42 million employees

Industry Trends

- Commerce anywhere, anytime
- Personal and in context

Scenarios

- Sales & Marketing: Demand Forecasting
- Customer & Channel: Personalized Offers

Healthcare

Industry Overview

- $2 trillion spent in U.S./year on healthcare
- 5,754 hospitals with >36 million patients
- Hospital readmissions cost ~$41 billion/year
- ACA new regulations for Medicare/Medicaid

Industry Trends

- Payments based on patient outcome
- Government monitoring hospital readmissions and number of hospital-acquired conditions

Scenarios

- Customer & Channel: Patient Outcomes
- Operations & Workforce: Operational Efficiency

Financial Services

Industry Overview

- 7.9% of US Economy
- $1.2 trillion revenue
- 12% revenue growth rate
- 5.9 million employees

Industry Trends

- Understanding customer behavior
- Cybersecurity
- Taking banking “mobile”
- Proactive regulatory compliance

Scenarios

- Finance & Risk: Risk & Compliance
- Sales & Marketing: Personalization

Big Data in Retail

Big Data in Retail

- Clickstream analytics
- Online recommendation engine
- 360° view of the customer
- Analyze brand sentiment
- Localized, personalized promotions
- Optimal store layout

Leading computer manufacturer in world

- Use clickstream to deliver custom website ecommerce experience
- Targeted ads for abandoned carts

SHINSEGAE

- Use unstructured data from website and social for data mining
- Combine w/sales data for 360 view

Leading clothing online retailer

- Use clickstream to understand who is viewing their site
- Building recommendation engine based on users' clickpaths

BlackBall™

- Gather social media sentiment from customer feedback
- Combined with POS data, can determine right product mix

Leading Multi-national Retailer

- Track weather information (temperature/forecast) to predict shelf space for different seasons
- Sentiment analysis on feedback

ziosk

- Gather data from table-side devices at restaurants
- Predict promotions/offers and content to upsell to guests

Ziosk turned to Microsoft gold partner, Artis Consulting to deploy a hybrid deployment consisting of the Analytics Platform System, Azure HDInsight, Power BI, and Azure Machine Learning.

"Until now, we haven't had the ability to optimize the guest experience based on their specific interactions with the devices. With Azure, we can close the loop."

Kevin Mowry
Ziosk
Chief Software Architect

Big Data in Healthcare

Big Data in Health

- Predictive Analysis of Patient Health & Clinical Decision Support
- Population, risk, and Care management
- Real-time quality measures to assist providers w/regulatory requirements
- Medical research data (eg. genomics)
- Recruit cohorts for pharmaceutical trials

ASCRIBE

- Scan handwritten notes and do natural language processing
- Analyze if symptoms might map to bigger outbreak

McKESSON

- Process large volumes of data from any healthcare provider EHR system
- Assist in showing compliance

NHS

- Process 6 years worth of data in a few hours without any infrastructure

CHR HANSEN

- Collect clinical trial data (from automated equipment, sensors)
- Find patterns on this data (chemical compositions, enzymes)

Beth Israel Deaconess Medical Center

- Store 7-30 years of data to meet audit requirements

Big Data in Financial Services

Big Data Financial Services

- New account risk screens
- Fraud prevention
- Trading risk
- Maximize deposit spread
- Insurance underwriting
- Accelerate loan processing

RBS
The Royal Bank of Scotland

- Actively monitor currencies used by UK manufacturers in supply chain to do risk analysis
- Monitor UK GDP to help customers stay on top of economic trends

DirectEdge

- Needed to handle increasing amounts of finance, compliance, and legal data from trading operations
- Trading data drives strategic decisions

Tangerine

- Track customer feedback on social media and on their blog posts/website to understand loyalty
- Predict at-risks clients to reach out to

Milliman

- Process data for actuaries to analyze results to understand risks for insurance companies
- Milliman's application understands relationships between people, process, and technology to manage risk

Tangerine partners with Microsoft to build a solution with Analytics Platform System for the data warehouse and uses PolyBase to query Azure HDInsight in the cloud.

"With pre-built integration using PolyBase to query both the relational data warehouse and Hadoop in the cloud, the solution will allow us to reap the benefits of both relational and non-relational data regardless of where it lives."

Kevin Mowry
Ziosk
Chief Software Architect

Tangerine

Agenda

Big Data As Part of Cortana Analytics

Business Scenarios For Big Data

What To Do Next

Attend big data sessions

Today – Part 2

- Hands-on Getting Started with query, storm, and HBase
  - 1:00
  - Baker

Tomorrow

- New services in the Cortana Analytics Suite
  - 10-12
  - Hood
  - Azure Data Lake, Apache Spark, Data Catalog

Learn More

Read documentation

http://azure.microsoft.com/en-us/documentation/services/hdinsight/

Learning Map

http://azure.microsoft.com/en-us/documentation/articles/hdinsight-learn-map/

Microsoft Virtual Academy

http://www.microsoftvirtualacademy.com/training-courses/getting-started-with-microsoft-big-data

Channel 9 Data Exposed Show

http://channel9.msdn.com/Shows/Data-Exposed

Try 30 day trial

http://azure.microsoft.com/en-us/pricing/free-trial/

Microsoft

© 2015 Microsoft Corporation. All rights reserved.

Session 2 – Big Data

NishantH@Microsoft.com

Agenda

Querying Data

Streaming

NoSQL

Big Data Applications

Agenda

Querying Data

Streaming

NoSQL

Big Data Applications

Managed Hadoop clusters in Azure

- Easy to deploy
- Integrated with Azure
- Automation with Azure Data Factory

Use cases: Analytics using various sources

- Aggregation Scenario
  - Hive allows data to be read from different sources
  - Transform data using Hive/Pig and write back to Azure Blob
  - Use Sqoop/ADF to move data to relational database

Use cases: Extract Transform Load

- ETL scenario
  - Extract data from variety of sources to Azure Blob
  - Transform data using Hive/Pig and write back to Azure Blob
  - Use Sqoop/ADF to move data to relational database for interactive usage

Use cases: Recommendation engines

- Recommendation scenario
  - Data from users is collected using Web logs and uploaded to Azure Blob
  - Recommendation tables are built using Mahout/Hive
  - Customers are shown appropriate titles based on their tastes

Hive

Overview

SQL for Hadoop!
PB Scale
If you know SQL, you know Hive!

Hive properties

Easier than Map Reduce
Schema on read
Supports user code
Reads multiple formats

Hive architecture

Hive compiles, Hadoop executes

Demo

```
DROP TABLE IF EXISTS weblogs;

CREATE EXTERNAL TABLE weblogs(s_date date, s_time string, s_sitename string, cs_method string, cs_uristem string,
                              cs_uriquery string, s_port int, cs_username string, c_ip string, cs_useragent string,
                              cs_cookie string, cs_referer string, cs_host string, sc_status int, sc_substatus int,
                              sc_win32status int, sc_bytes int, cs_bytes int, s_timetaken int )
ROW FORMAT DELIMITED FIELDS TERMINATED BY ' '
STORED AS TEXTFILE LOCATION 'wasb://rashimg6@rashimgstorage.blob.core.windows.net/HdiSamples/WebsiteLogSampleData/SampleLog/'
TBLPROPERTIES ('skip.header.line.count'='2');

select * from weblogs;


DROP TABLE IF EXISTS HttpStatus;

CREATE EXTERNAL TABLE HttpStatus(code int, status string)
ROW FORMAT DELIMITED FIELDS TERMINATED BY ','
STORED AS TEXTFILE LOCATION 'wasb://rashimg6@rashimgstorage.blob.core.windows.net/HdiSamples/WebsiteLogSampleData/Status/';

SELECT * from HttpStatus;

-- Join

SELECT weblogs.cs_referer, weblogs.sc_status, HttpStatus.status from weblogs INNER JOIN HttpStatus WHERE weblogs.sc_status=HttpStatus.code;
```

Hive Optimizations

Scale up/Scale out

Tez

Partitioning

ORCFile

Vectorization

Agenda

Querying Data

Streaming

NoSQL

Big Data Applications

Streaming Analytics

Storm Essentials

- What is Storm
- Storm Use Cases
- Introduction to Storm
- Storm Topologies
- Storm on HDInsight

What is Apache Storm?

Streaming data analysis

Easy to program

A distributed real time processing platform

Fast

Clocked at 1M+ messages per second per node

Scalable

Thousands of workers per cluster

Fault Tolerant

Failure is expected, and embraced

Reliable

Guaranteed message delivery Exactly-once semantics

Storm Essentials

- What is Storm
- Introduction to Storm
- Storm Topologies
- Storm on HDInsight

Storm introduction

Tuple

- Core unit of data
- Immutable set of key/value pair

Stream

- Unbounded sequence of Tuples

Spout

- Source of streams
- Wraps a streaming data source and emits Tuples

Using Bolts

Core functions of a streaming computation | Receive tuples and do ***stuff***

Storm Essentials

- What is Storm
- Introduction to Storm
- Storm Topologies
- Storm on HDInsight

Topologies

Demo

Cloud gateways
Data Generator
Counter Bolt
Aggregate Writer Bolt
HBASE
Live dashboard

Storm Essentials

- What is Storm
- Introduction to Storm
- Storm Topologies
- Storm on HDInsight

Apache Storm on HDInsight

- Managed services
- Open source platform
- Analyse millions of messages per second
- Scale-up and scale-down
- Speed
- Event Hub
- Visual Studio
- Azure
- HBase, SQL Database, DocumentDB

Visual Studio tooling

Support for authoring Storm Topologies

- Create Storm projects from available template
- Submit a topology with C# bolts/spouts
- Submit Topologies containing Java spouts/bolts
- Monitor Topologies within VS
- Troubleshoot Topologies

In essence, you never need to leave Visual Studio for Storm Projects

Technology comparison

		Storm on HDInsight	Azure Stream Analytics
Management & Operations	Service	Managed Cluster	Managed Service
	Price	Link to the pricing page.	Link to the pricing page.
	Microsoft Supported	Yes	Yes
	Open Source	Yes	No
Development Experience	SQL DSL	No	Yes
	Extensible	Yes	No
	Temporal Operators	No. Customer write custom code	Yes
	Authorig/Debugging Tools	via Visual Studio	Interactive authoring and debugging via azure portal
Input/Output	Data Ingress	No restriction	Event Hub, Azure Blobs
	Data egress	No restriction	Support to write data to Event Hubs, Blob store, azure table, azure sql db, Powerbi
	Supports Multiple Inputs	Yes	Yes
	Generate Multiple Outputs	Yes	Yes
	Data Format	No restriction	Avro, JSON,CSV
Performance	Scalability	Yes	Yes
	Elastic Scale	Yes	Yes

Agenda

Querying Data
Streaming
NoSQL
Big Data Applications

What is HBase?

- Distributed, non-relational database
  - Columnar, schema-free data model
  - NoSQL on top of Hadoop
- Large scale
  - Linear scalability
  - Billions of rows X millions of columns
  - Transparent auto-sharding of big tables
- Low latency
  - Low latency random reads, in-memory caching
  - High throughput writes
- Open Source
  - Started in 2006, modeled after Google's BigTable

Use cases: key value store

- Key value store
  - Message systems
  - Content management systems
- Examples
  - Facebook Messages
  - Twitter-like messages
  - Webtable – web crawler/indexer

Use cases: sensor data

- Sensor data
  - Social analytics
  - Time series databases
  - Interactive dashboards with trends, counters, etc
  - Audit log systems
- Examples
  - Bloomberg trader terminal
  - OpenTSDB

Use cases: sensor data

- Architecture

HBase Demo: Twitter Sentiment App

HBase Resources

- HBase Book

  HBase: The Definitive Guide

  Online HBase Book
- C# HBase SDK

  https://github.com/hdinsight/hbase-sdk-for-net
- Source Code: Tweet Sentiment App

  https://github.com/maxluk/tweet-sentiment
- Documentation

  Get started using HBase in HDInsight

  Tutorial: Building Tweet Sentiment App

Big Data Applications

Security, Identity, Data Prep and Analytics… to start.

Data Analytics - Microsc

ms.portal.azure.com/?r=1#blade/HubsExtension/GalleryFeaturedMenuItemBlade/selectedMenuItemId/DataAnalytics_MP

Add to the favorites bar by selecting ☆, or by importing your favorites.

Internal | Microsoft Azure > Create > Data + Analytics > Marketplace > Data Analytics | Report bug | Nino Bice MICROSOFT

NEW
HOME
BROWSE ALL
NOTIFICATIONS
ACTIVE
BILLING

Marketplace

- Everything
- Compute
- Web + Mobile
- Data + Storage
- Data Analytics
- Networking
- Media + CDN
- Hybrid Integration
- Security + Identity
- Developer Services
- Management
- Container Apps

Data Analytics

Filter

Search Data Analytics

Search
Microsoft

Data Factory
Microsoft

MySQL Database
ClearDB

Data analytics — More

Hortonworks Data Platform
Hortonworks

HDInsight
Microsoft

Cloudera Enterprise
Cloudera

Data insights — More

RedPoint Interaction™ -

ms.portal.azure.com/?r=1#blade/HubsExtension/GalleryFeaturedMenuItemBlade/selectedMenuItemId/DataAnalytics_MP

Add to the favorites bar by selecting ☆, or by importing your favorites.

Internal | Microsoft Azure > Create > Data + Analytics > Marketplace > Data Analytics > RedPoint Interaction™ | Report bug | Nino Bice MICROSOFT

NEW
HOME
BROWSE ALL
NOTIFICATIONS
ACTIVE
BILLING

Microsoft | Cloudera

More

g Search API
rosoft

D&B Business Insight
D&B

Embarke Email Analytics
Embarke

Worldwide Historical Weather
Weather Trends Int...

RedPoint Interaction™
RedPoint Global Inc.

More

es and Use Tax
es Complete

IP Check

Phone Number Validation

Verify

RedPoint Interaction™
RedPoint Global Inc.

Bring Your Own License enabled.
This virtual machine uses Windows Server, SQL Server Standard Edition. Additional fees apply.

Campaign Management to Drive Rich, Omnichannel Customer Interactions

Execute across the complete lifecycle of customer engagement with a campaign management solution that makes every interaction more personalized and engaging. With RedPoint Interaction™ you can engage customers in a continuous dialog, across the full spectrum of channels – digital, traditional, inbound, outbound – all from a single visual interface.

RedPoint Interaction is database and data model-agnostic. It scales to meet growing needs rapidly and cost-effectively. RedPoint Interaction delivers pre-built integrations with leading email services, mobile messaging services, social media, and more. It facilitates teamwork across fragmented organizations with integrated collaboration tools, and allows you to plan, budget and track results.

RedPoint was rated a Leader in The Forrester Wave™: Cross-Channel Campaign Management, Q3 2014 in its first appearance. It was ranked highest in cross-channel integration and customer satisfaction.

If you're interested in RedPoint Interaction, download the following checklist & order form at RedPoint Interaction Client\Environmental Details. Complete each of the required setup steps in order to customize RedPoint Interaction for your virtual environment.

Select a deployment model
Classic

Create

Want to deploy programmatically? Get started →

NEW

HOME

0 NOTIFICATIONS

BROWSE

ACTIVE

BILLING

{ISV App Name}

1 BASICS
Configure your app

2 HDINSIGHT CLUSTER
Configure your cluster

3 VIRTUAL NETWORK
Configure your virtual network

4 SUMMARY
See a deployment summary

Create

Basics

* Name

* User name

* Authentication type

Password | SSH public key

* Password

* Subscription
HDInsight_Telemetry

* Resource group

Select existing

* Location
East US

OK

NEW

HOME

0 NOTIFICATIONS

BROWSE

ACTIVE

BILLING

New HDInsight Cluster

1 BASICS
Configure your app

2 HDINSIGHT CLUSTER
Configure your cluster

3 VIRTUAL NETWORK
Configure your virtual network

4 SUMMARY
See a deployment summary

Create

New HDInsight Cluster

* Cluster Name
mycluster001
.azurehdinsight.net

Cluster Type
Storm

Cluster Operating System
Linux (Ubuntu 12.04 LTS)

* Subscription
HDInsight_Telemetry

* Resource Group
_hditelemetry(default)

Create New

* Credentials
Configured

* Data Source
abc12341 (East US)

* Node Pricing Tiers
D3/A3/A3

Optional Configuration

Pin to Startboard

Create

NEW

HOME

0 NOTIFICATIONS

BROWSE

ACTIVE

BILLING

New HDInsight Cluster

1 **BASICS**
Configure your app

2 **HDINSIGHT CLUSTER**
Configure your cluster

3 **VIRTUAL NETWORK**
Configure your virtual network

4 **SUMMARY**
See a deployment summary

Create

New Virtual Network

* Name

hditelem

* Address space

10.0.0.0/16

10.0.0.0 - 10.0.255.255 (65536 addresses)

* Subnet name

default

* Subnet address range

10.0.0.0/24

10.0.0.0 - 10.0.0.255 (256 addresses)

OK

NEW

HOME

0 NOTIFICATIONS

BROWSE

ACTIVE

BILLING

New HDInsight Cluster

1 **BASICS**
Configure your app

2 **HDINSIGHT CLUSTER**
Configure your cluster

3 **VIRTUAL NETWORK**
Configure your virtual network

4 **SUMMARY**
See a deployment summary

Create

Summary

Subscription	HDInsight_Telemetry
Resource group	(new) hditelem
Location	East US
Computer name	vmname
User name	admina
Size	Standard D1
Disk type	Standard
Storage account	(new) hditelem3270
Virtual network	(new) hditelem
Subnet	(new) default (10.0.0.0/24)
Public IP address	(new) vmname
Network security group	(new) vmname
Availability set	None
Diagnostics	Enabled
Diagnostics storage account	(new) hditelem3270

OK

HDI Application Demo

Microsoft

© 2015 Microsoft Corporation. All rights reserved.