Congress on Harnessing Big Data for the Environment: Private Sector Capabilities

Kristin M. Tolle, Ph.D.
Director, Program Management Advanced Analytics Training and Ecosystem Development
December 6th, 2016
AGENDA

• General Azure Overview
• Case Study: Predictive Analytics for Flooding
• Case Study: Energy Demand Forecasting
• Case Study: Deep Learning for Massive Datasets – Galaxy Classification
The Cloud is Ubiquitous

34

Azure Regions
The Cloud is Secure

More certifications than any other cloud provider

Industry leader for customer advocacy and privacy protection

Unique data residency guarantees
Choice + Flexibility

Management
- puppet labs
- CHEF
- ANSIBLE

Applications
- apprenda®
- WordPress
- Joomla!
- Drupal
- CLOUD FOUNDRY

App Frameworks
- nodeJS
- python
- Java
- PHP
- Ruby

Databases & Middleware
- Hortonworks
- Cloudera
- MongoDB
- MySQL

Infrastructure
- FreeBSD
- Red Hat
- Linux
- Docker
Hybrid Management and Security

Log analytics
Automation
Backup
DR and data protection
Security
CORTANA INTELLIGENCE SUITE
TRANSFORM DATA INTO INTELLIGENT ACTION

Data Sources
- Data Factory
- Data Catalog
- Event Hub

Big Data Stores
- Data Lake Store
- Blob Store
- SQL Data Warehouse

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

Intelligence
- Cognitive Services
- Bot Framework
- Cortana

Business Scenarios
- Recommendations, customer churn, forecasting, etc.

Dashboards & Visualizations
- Power BI

Apps
- Web
- Mobile
- Bots

People

Automated Systems

Data Sources → Information Management → Big Data Stores → Machine Learning and Analytics → Intelligence → Business Scenarios → Dashboards & Visualizations → Apps → People → Action

Data → Intelligence → Action
FLOOD FORECASTING
**National Flood Interoperability Experiment:**

**Close the gap between national flood forecasting and local emergency response**

- Improve information flow
- National forecasting at stream reach scale
AUTHORITATIVE DATA SOURCES

Past

USGS
Monitoring Statistics

USACE
Reservoirs, Levees

NWS
Forecasts

Present

Future

FEMA
Flood Mapping
CHALLENGE: INTEGRATING SPACE AND TIME

Spatial Aggregation

Temporal Aggregation

Database

Web Solution

Spatial Services

Temporal Services

Servers

Space

Time
BRINGING THE PIECES TOGETHER...
ADVANCED FLOOD MODELING ON CONSISTENT SPATIAL DOMAIN

Process Geospatial Data

Model Stream Flows

Create Flood Data

Azure ML

Stream Flow

Flood Map

GeoProcessing

Cortana Advanced Analytics

Local Alerts

Impact Assessment

Planning Mitigation Response

USACE Reservoir Releases

Localize Forecast (RAPID)

NCAR Gridded Forecasts

NWS Basin Forecasts

NCEP

NHDPlus
CALCULATING SOIL MOISTURE MODELS

http://statsnldas.azurewebsites.net

Source: Gonzalo Espinoza Davalos
MACHINE LEARNING TO ESTIMATE MISSING VALUES

Source: Tim Petty and Prashant Dhingra
Prediction results for DischargeRate_13185000

variable
- Actual
- Scored.Labels
NEXT STEP... SCALE TO THE US

Calculate ~3 million river reaches
ENERGY FORECASTING
**ARCHITECTURE**

- **Azure WebJob**: Runs jobs to get data from public source.
- **Event Hub**: Stores streaming data. Stream Analytics processes events as they arrive in the EventHub.
- **Stream Analytics**: Processes events as they arrive in the EventHub.
- **aml Model**: Web Service.
- **Azure Data Factory Pipeline invokes AML Web Service.**
- **Power BI Dashboard**: BES endpoint.
- **Real-time data stats.**
- **Hourly Prediction Updates.**
- **Real Time Batch.**
- **Send to Azure SQL for batch predictions.**
- **Azure SQL**: Contains historical arrival/departure consumption & weather data.
- **Azure Data Factory Pipeline invokes AML Web Service.**
- **AML Model Web Service BES endpoint.**

**BaneDK Data**

**Azure Services**

**Weather and Train Data sources**
energyforecast_5minsdata_asa

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

query

Query can't be edited while a job is running.

Power BI is designed for cases where Azure Stream Analytics does a significant data load reduction by aggregating on a hopping or tumbling window.

```
15  SELECT a.TimeStamp as Time, a.TimeZone, a.Name as Region, a.PTID, b.Latitude, b.Longitude, a.Load as Demand 
16  INTO outputPBI  
17  FROM 
18  (SELECT TimeStamp, TimeZone, Name, PTID, Load FROM InputEventHub) a 
19  join 
20  (SELECT PTID, Name, Latitude, Longitude FROM InputBlobRefData) b 
21  on a.PTID=b.PTID 
22  ;
```

Missing some language constructs? Let us know! (Powered by UserVoice - Privacy Policy)
AZURE MACHINE LEARNING

EnergyDemandForecast-24hr-withT

Properties
- Reader
  - Data source: Azure SQL Database
  - Database server name: ghdh7weg5iww.database.windows.net
  - Database name: EnergyForecastSQLDatabase
  - Server user account name: energydemouser
  - Server user account password: ************
  - Database query:
    ```sql
    1 select b.time as TimeStamp, b.PTID, a.HourAvg
    2 (select * from DemandHistoryHourly)
    3 where convert(varchar(10),TimeStamp,110)=convert(v
    4 and PTID=01701) a
    5 right join
    ```
- Quick Help
  - Load data from sources such as the Web, Azure SQL database, Azure table, Hive table, or Windows Azure BLOB storage
Azure Machine Learning

EnergyDemandForecast-24hr-withTemp - Demo ➤ Reader ➤ Results dataset

<table>
<thead>
<tr>
<th>TimeStamp</th>
<th>PTID</th>
<th>HourAvgLoad</th>
<th>temperature</th>
</tr>
</thead>
<tbody>
<tr>
<td>2012-09-12T00:00:00</td>
<td>61761</td>
<td>5090.6</td>
<td>59.61</td>
</tr>
<tr>
<td>2012-09-12T01:00:00</td>
<td>61761</td>
<td>4797.5</td>
<td>60.52</td>
</tr>
<tr>
<td>2012-09-12T02:00:00</td>
<td>61761</td>
<td>4624.4</td>
<td>60.14</td>
</tr>
<tr>
<td>2012-09-12T03:00:00</td>
<td>61761</td>
<td>4544.7</td>
<td>59.59</td>
</tr>
<tr>
<td>2012-09-12T04:00:00</td>
<td>61761</td>
<td>4546.1</td>
<td>58.47</td>
</tr>
<tr>
<td>2012-09-12T05:00:00</td>
<td>61761</td>
<td>4786.3</td>
<td>57.85</td>
</tr>
<tr>
<td>2012-09-12T06:00:00</td>
<td>61761</td>
<td>5383.9</td>
<td>57.79</td>
</tr>
<tr>
<td>2012-09-12T07:00:00</td>
<td>61761</td>
<td>6037.4</td>
<td>59.12</td>
</tr>
<tr>
<td>2012-09-12T08:00:00</td>
<td>61761</td>
<td>6573.9</td>
<td>63.09</td>
</tr>
<tr>
<td>2012-09-12T09:00:00</td>
<td>61761</td>
<td>6978.5</td>
<td>67.45</td>
</tr>
<tr>
<td>2012-09-12T10:00:00</td>
<td>61761</td>
<td>7214.9</td>
<td>69.9</td>
</tr>
<tr>
<td>2012-09-12T11:00:00</td>
<td>61761</td>
<td>7395.9</td>
<td>71.72</td>
</tr>
<tr>
<td>2012-09-12T12:00:00</td>
<td>61761</td>
<td>7524.2</td>
<td>72.1</td>
</tr>
</tbody>
</table>

Statistics

- Mean: 5604.3624
- Median: 5651.45
- Min: 2859.6001
- Max: 10610.9004
- Standard Deviation: 946.7895
- Unique Values: 7099
- Missing Values: 48
- Feature Type: Numeric Feature

Visualizations

HourAvgLoad Histogram

compare to: None

![Histogram Chart]
AZURE MACHINE LEARNING
energydemandforecast-24hr-withtemp

Default Endpoint

**API Help Page**

**Test**

**Last Updated**

9/2/2015 7:42:40 AM

9/2/2015 7:42:40 AM

Additional endpoints

Number of additional endpoints created for this web service: 1

Manage endpoints in Azure management portal
AZURE DATA FACTORY
AZURE DATA FACTORY
### Azure SQL Database

```sql
SELECT Timestamp, PTID, Name, Demand, Forecast, Hi95, Lo95, APE FROM v_24hrsV2ForecastPointErrorOngoing;
```

<table>
<thead>
<tr>
<th>Timestamp</th>
<th>PTID</th>
<th>Name</th>
<th>Demand</th>
<th>Forecast</th>
<th>Hi95</th>
<th>Lo95</th>
<th>APE</th>
</tr>
</thead>
<tbody>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61752</td>
<td>WEST</td>
<td>1977.308</td>
<td>1945.6005859375</td>
<td>2087.82565293257</td>
<td>1823.575519458243</td>
<td>1.80358214631213</td>
</tr>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61753</td>
<td>GENES</td>
<td>1274.492</td>
<td>1236.68481445313</td>
<td>1339.22356717661</td>
<td>1134.14636172964</td>
<td>2.96642848184266</td>
</tr>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61754</td>
<td>CENTR</td>
<td>1906.825</td>
<td>1903.62426757813</td>
<td>2027.50057235469</td>
<td>1779.74796280156</td>
<td>0.16784086017595</td>
</tr>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61755</td>
<td>NORTH</td>
<td>466.2583</td>
<td>480.52142339844</td>
<td>501.98888504588</td>
<td>459.0561781751</td>
<td>3.0590535112562</td>
</tr>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61756</td>
<td>MHK VL</td>
<td>980.6583</td>
<td>996.000024414063</td>
<td>1078.255258599</td>
<td>915.544752909122</td>
<td>1.65620367476263</td>
</tr>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61757</td>
<td>CAPITL</td>
<td>1599.533</td>
<td>1576.87219238281</td>
<td>1675.10122652832</td>
<td>1478.64371818228</td>
<td>1.47332832288918</td>
</tr>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61758</td>
<td>HUD VL</td>
<td>1329.575</td>
<td>1328.86810839844</td>
<td>1373.3138754749</td>
<td>1192.45834132198</td>
<td>3.51568150702596</td>
</tr>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61759</td>
<td>MILLWD</td>
<td>355.5917</td>
<td>334.907653803894</td>
<td>368.824894346224</td>
<td>300.990413270963</td>
<td>5.8167900428074</td>
</tr>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61760</td>
<td>DUNWOD</td>
<td>923.9</td>
<td>858.163269042989</td>
<td>922.319857476789</td>
<td>794.00668060948</td>
<td>7.11537312932959</td>
</tr>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61761</td>
<td>N.Y.C.</td>
<td>7634.192</td>
<td>7665.3486328125</td>
<td>8111.29290089669</td>
<td>7219.40436472831</td>
<td>0.408120975351843</td>
</tr>
<tr>
<td>2015-09-09 00:00:00.000</td>
<td>61762</td>
<td>LONGIL</td>
<td>2089.088</td>
<td>2047.04565429668</td>
<td>3173.26100377251</td>
<td>2720.83030482124</td>
<td>1.40558452800465</td>
</tr>
</tbody>
</table>
Power BI CAW Energy Forecast Dashboard

Insights

Average Demand by Region

Last 24hrs Peak Demand

Last 24hrs Off-Peak Demand

Total Demand by Region

Forecast

Average Temperature by Time

Demand vs Forecast with Confidence Interval

Error Rate by Time (%)
I believe over the next decade computing will become even more ubiquitous and intelligence will become ambient... This will be made possible by an ever-growing network of connected devices, incredible computing capacity from the cloud, insights from big data, and intelligence from machine learning.

Satya Nadella,
CEO, Microsoft
Deep Learning’s First Major Success: Speech Recognition

NIST “Switchboard” speech recognition challenge
Experience WorldWide Telescope

Immerse yourself in a seamless web experience...

From web to desktop to full dome projection, WorldWide Telescope takes you on a journey through the universe, bringing together telescopes in the world and combining live data from astronomers and educators to import your own data and visualize it. A web-based version of WorldWide Telescope offers guided explorations of the universe for the first time using the power of Microsoft Silverlight.

Worldwide Telescope Layerscape: Visualize, Explore, and Discover your data. Visualize your universe of data, and share it through click-and-drag, zoom and pan, image, vector and tabular data as well as your data. You can appreciate the Virtual Observatory clients on your own... Download the update today. Just click on the link...

http://www.worldwidetelescope.org/
CLASSIFYING GALAXY STRUCTURE USING NEURAL NETWORKS WITH R
100 Billion Galaxies in observable universe

2 trillion Hubble ultra deep

200 billion Hubble deep field

GALAXY SHAPE TELLS US ABOUT EVOLUTION

**GALAXY SHAPE TELLS US ABOUT EVOLUTION**

- **Spiral galaxies**
- **Elliptical galaxies**

- Collisions and other events

- Forming
- Ancient
ARTIFICIAL NEURAL NETWORKS
How do computers “see”?

- Convolutional Deep Belief Networks for Scalable Unsupervised Learning of Hierarchical Representation
- Honglak Lee, Roger Grosse, Rajesh Ranganath, Andrew Y. Ng
WHAT COMPUTERS “SEE”

A two-dimensional array of pixels

Neural network

Spiral
Elliptical
THE COMPLICATION

rotation  scaling  translation

Neural network
THE SOLUTION

Match pieces of the image
**THEN REPEAT ACROSS THE ENTIRE IMAGE**

Convolution

Matches specific shape (kernel) across entire image

Automatic feature generation
DEEP STACKING

Layers can be repeated several (or many) times
Microsoft R Server – R for Big Data

The All-Inclusive Big Data Big Analytics Platform

High-performance open source R:
• Scale & Performance
  • Scales from workstations to large clusters
  • Scales to large data sizes
  • Growing portfolio of Parallelized algorithms
• Write Once Deploy Anywhere - multiple platforms
• IDE for data scientists and developers
TRAINING ON AZURE

- Skyserver database
- SDSS
- Storage blob Images
- SQL Server
- SQL2016 R Services
- Data Science Virtual machine
- Azure Web
- Azure N Series GPU VM
- Train model
- Azure storage
<table>
<thead>
<tr>
<th>100K * 3</th>
<th>Training images, augmented with rotation</th>
</tr>
</thead>
<tbody>
<tr>
<td>8</td>
<td>Layers in deep network</td>
</tr>
<tr>
<td>176K</td>
<td>Weights to compute in network</td>
</tr>
<tr>
<td>2.5B</td>
<td>Weight updates per second</td>
</tr>
<tr>
<td>1.8 hours</td>
<td>Computing time on Azure N series GPU</td>
</tr>
<tr>
<td>88%</td>
<td>Overall accuracy - training data</td>
</tr>
<tr>
<td>55%</td>
<td>Overall accuracy - test data (91% micro-average)</td>
</tr>
</tbody>
</table>

The technique works, but has scope for improvement!
“Big Data for Healthy Cities: Using Location-aware Technologies, Open Data and 3D Urban Models to Design Healthier Built Environments”

H. Miller and K. Tolle
TRANSFORMING SCIENCE TO BEING SERVICE-CENTRIC

Gather data from devices

Transform data into information and insight

Monitor processes and assets remotely

Connectivity, open data access and interoperability transforms changes scientific focus from reactive to proactive

Take corrective action anytime, anywhere

Enable inclusive global cross-functional collaboration