



# **Operational Machine Learning**

Using Microsoft Technologies for Applied Data Science

Khalid M. Salama, Ph.D.

Business Insights & Analytics Hitachi Consulting UK

# Outline



- Introduction to Data Science
- From Experimental Data Science to Operational Machine Learning
- MS Technologies for Data Science & Advanced Analytics
- Demos & Screenshots
- Concluding Remarks



# Introduction to Data Science and Machine Learning



### What?

"Data mining, an interdisciplinary subfield of computer science, is the computational process of automatic discovering interesting and useful patterns in large data sets"

# **Other Related Technologies:**

- Visualization
- Big Data
- High Performance Computing
- Cloud Computing
- Others..

4



# **Data Science and Machine Learning**



### Why?

The objective of data science is to provide you with actionable insights to support decision making....



# **Data Science and Machine Learning**



### How?

### **Classification Learning**

Build a model that can predict the target class of an input case

Build a model that can estimate the response

### **Regression Modeling**

### **Cluster Analysis**

value given an input case

Discover natural groupings within the data points

### **Association Rule Discovery**

Extract frequent patterns present in the data

IF .. AND .. AND .. THEN A ELSE IF .. AND .. THEN C ELSE IF .. AND .. THEN B .. ELSE C

### **Time Series Analysis**

Analysis of temporal data to forecast future values

### **Probabilistic Modeling**

Compute the probability of an event to occur given a set of conditions

### **Similarity Analysis**

Identify similar cases to a given input case based on the input features

### **Collaborative Filtering**

Filtering of information using techniques involving collaboration viewpoints





# £ <mark>- ■</mark> ×



# From Experimental Data Science to Operational Machine Learning

# **Data Science Activities**



### **Data Analysis & Experimentation**

- Interactive
- Easy to perform
- Rich Visualizations



HITACHI

**Inspire the Next** 

# **Data Science Activities**



# Experimentation vs. Operationalization

## **Operational ML Pipelines**

- Pipelined (ETL Integration)
- Scalable
- Apps Integration





# Microsoft Advanced Analytics Technologies

# **Microsoft Advanced Analytics**



### Cortana Intelligence Suite

### https://gallery.cortanaintelligence.com/





### Data Science, Machine Learning, & Intelligence





# **Microsoft Azure Machine Learning**





## MS Cloud-native Data Science

- Cloud-based Machine Learning Services
- Interactive Data Science Studio
- Rich built-in functionality
- Imports data from everywhere
- Easy to develop and productionize Web Services
- Extensible via R and Python scripts



- Only Cloud-based (Data Regulations)
- Scalability Maximum dataset size = 10GB
- Microsoft R Open is not supported, yet
- No Source Control







### **Real-time Predictions**





### **Built-in Features**

💑 Saved Datasets 🔒 Data Format Conversions 🔂 Data Input and Output 🐫 Data Transformation Seature Selection Machine Learning DenCV Library Modules Python Language Modules R Language Modules **Statistical Functions** Text Analytics Web Service Deprecated

- Data Format Conversions Data Transformation Convert to ARFF Convert to CSV Convert to Dataset Convert to SVMLight Convert to TSV bata Input and Output Enter Data Manually Export Data Import Data Unpack Zipped Datasets Filter Apply Filter FIR Filter **IIR Filter** Median Filter Moving Average Filter Threshold Filter User Defined Filter
  - Learning with Counts Manipulation Sample and Split Scale and Reduce Manipulation Add Columns Add Rows Apply SQL Transformat... Clean Missing Data Convert to Indicator Va. Edit Metadata Group Categorical Valu... Join Data Remove Duplicate Rows Select Columns in Data... Select Columns Transfo.. SMOTE

Filter

- ▲ ∑<sub>II</sub> Statistical Functions Apply Math Operation Compute Elementary Statist... Compute Linear Correlation Evaluate Probability Function Replace Discrete Values Summarize Data Test Hypothesis using t-Test Python Language Modules Execute Python Script R Language Modules Create R Model Execute R Script ▲ ● Feature Selection Filter Based Feature Selectio.. Fisher Linear Discriminant A... Permutation Feature Import...
- Text Analytics Feature Hashing Named Entity Recognition Score Vowpal Wabbit Versio.. Score Vowpal Wabbit Versio.. Score Vowpal Wabbit Versio... Train Vowpal Wabbit Versio... Train Vowpal Wabbit Versio... Train Vowpal Wabbit Versio... Sample and Split Partition and Sample Split Data Scale and Reduce Clip Values Group Data into Bins Normalize Data Principal Component A...
- Machine Learning Evaluate Cross Validate Model Evaluate Model Evaluate Recommend... Initialize Model Anomaly Detection Classification Clustering Regression Score . Apply Transformation Assign Data to Clusters Score Matchbox Reco... Score Model Train Sweep Clustering Train Anomaly Detect... Train Clustering Model Train Matchbox Reco... Train Model Tune Model Hyperpa...





### Algorithms Cheat Sheet







### ML Studio





### Web Service

Sampl	le Code		
Rec	quest-Response	Batch	
C#	Python Pyt	hon 3+	R
// This // Inst // Tool	; code requires t ructions for doi .s -> Nuget Packa	he Nuget ing this ing Manage	package Microsoft.AspNet.WebApi.Client to be installed. n Visual Studio: r -> Package Manager Console
// Inst	all-Package Micr	osoft.Aspl	Net.WebApi.Client
using S	vstem;		
using S	ystem.Collection	s.Generic	
using S	ystem.IO;		
using S	ystem.Net.Http;		
using S	ystem.Net.Http.F	ormatting	, ,
using S	ystem.Net.Http.H	leaders;	
using S	ystem.Text;		
using S	ystem.Threading.	Tasks;	
namespa	ce CallRequestRe	sponseSer	vice
{			
cla	iss Program		
{			
	static void Mai	in(string[	] args)
	{		
	ToyokeReque		oSonvico() Wait();

Web service consu	umption options
х	
Excel 2013 or later Ex	cel 2010 or earlier
Basic consumption	n info
Want to see how to con	sume this information? Check out this easy tutorial.
Primary Key	hvNUKjLjUsd5sP+8P6vxwvQSlhcnpaY3MeploiicQQWI/AyDABdUz35THui0dgbRgn6C2UKXiAmAbj/R
Secondary Key	2PTJNK4unpg8zovCib0PfTsh4ZqKWmkLiX+UK8pWiGbrwvqR+NO7kzqdnzKfvOPgBIJJRf/dQQEXq/x
Request-Response	https://europewest.services.azureml.net/subscriptions/030b516729344453b3238d28287ae375/sen fc/execute?api-version=2.0&format=swagger
	Documentation
Batch Requests	https://europewest.services.azureml.net/subscriptions/030b516729344453b3238d28287ae375/serv fc/jobs?api-version=2.0

input	9	input	9
output	1	output	1
output	1		





### **Stream Analytics Integration**

(fx) sa02-sqlbits-estimation - F	unctions				
Search (Ctrl+/)	+ Add				
Overview	NAME	PARAMETERS	OUTPUT TYPE	FUNCTION TYPE	
Activity log	aml-EstimateOutput	2	record	Azure ML	
Access control (IAM)				sa	02-sqlbits-estimation
🖉 Tags				Que	y .
X Diagnose and solve problems				H	Save 🦿 Discard 🤀 Test
SETTINGS				-	E Inputs (1) Need help with your query? Check out some of the most common Stream Analytic
Locks					➢ input-eventhub-regdata 1 With estimation
					3 (
JOB TOPOLOGY				-	_, Outputs (2) 4 5 SELECT
					ind) output-powerbi 6 input,
III Functions					⇒ estimation 7 [aml-EstimateOutput](Input,0) as result 8 FROM
<> Query					9 [input-eventhub-regdata]
					10)
					12 SELECT
					13 Input,
					14 CAST(result.[Scored Labels] AS float) AS Output
					16 [output-powerbi]
					17 FROM
					18 estimation





### AzureML R Library

```
library(AzureML)
ws <- workspace(id = "1f2b56bcf5fe4f3f9c32ee437]</pre>
auth = "VsvDbAo+noBzPxphaoZZfdyY57T6QHDpdyM4P7W
api endpoint = "https://europewest.studio.azure
management endpoint = "https://europewest.management
head(experiments(ws))
data file = "C:/Master/data.csv"
data = read.csv(data file, header = TRUE)
model = lm(output \sim input, data = data)
test = data.frame(input = c(1, 10, 100))
predict(model, test)
predictOutput = function(input) {
    data = data.frame(c(input))
    colnames(data) = c("input")
    output = predict(model, data)
    return(output)
```



# **Microsoft R Server**

R in Microsoft World

### Microsoft R Open (MRO)

Microsoft R Server

- Based on latest Open Source R (3.2.2.) Built, tested, and distributed by Microsoft
- More efficient and multi-threaded computation
- Enhanced by Intel Math Kernel Library (MKL) to speed up linear algebra functions
- Compatible with all R-related software







Comparison

	CRAN	MRO	MRS
Data size	In-memory	In-memory	In-memory & disk
Efficiency	Single threaded	Multi-threaded	Multi-threaded, parallel processing 1:N servers
Support	Community	Community	Community + Commercial
Functionality	7500+ innovative analytic packages	7500+ innovative analytic packages	7500+ innovative packages + commercial parallel high- speed functions
Licence	Open Source	Open Source	Commercial license.

# Microsoft R Server



### **Components and Compute Contexts**

Scale & Deploy

Installed on Windows or Linux

**MS R Client** 

RStudio | RTVS

- ScaleR Optimized for parallel execution on Big Data, to eliminate memory limitations.
- ConnectR Provides access to local file systems, hdfs, hive, sqlserver, Teradata, etc.
- DistributeR Adaptable parallel execution framework to enable running on different (distributed) compute contexts.
- **Operationalization (msrdeploy)** Deploy the model as a Web API.







### Microsoft R Server – ScaleR Example

#### **Check Environment**

Revo.version	titanic xdf = "data/titanic xdf"
Revo.home()	rxImport(titanic_csv, titanic_xdf, colClasses = col_classes, overwrite = TRUE)
<pre>rxGetComputeContext()</pre>	<pre>titanic_xdata &lt;- RxXdfData(titanic_xdf)</pre>
<pre>#rxSetComputeContext()</pre>	<pre>rxGetInfo(titanic_xdata, getVarInfo = TRUE, numRows = 1)</pre>
	rxSummarv( ~ Survived, titanic xdata)

#### Prepare Data – Process XDF

```
rxDataStep(titanic_xdata, titanic_xdata,
transforms = list(
   Survived = factor(Survived, levels = 0:1, labels = c('No', 'Yes')),
   FareToAgeRatio = Fare/Age
   ),
   overwrite = TRUE)
```

```
prepare_data <- function(data) {
    age_mean = mean(data$Age, na.rm = TRUE)
    data$Age[is.na(data$Age)] <- age_mean
    return(data)
}
rxDataStep(titanic_xdata, titanic_xdata,
        transformFunc = prepare_data,
        overwrite = TRUE)</pre>
```

#### **Build Predictive Model**

#### **Perform Prediction**

```
test_data = data.frame(Age = c(30,20), Sex = c("male","female"))
predictions = rxPredict(rx_decision_tree, test_data)
head(predictions)
```



# Microsoft R Server – ScaleR Functionality

### Data Preparation

 Data import – Delimited, Fixed, SAS, SPSS, OBDC

Microsoft R Server

- Variable creation & transformation
- Recode variables
- Factor variables
- Missing value handling
- Sort, Merge, Split
- Aggregate by category (means, sums)

### **Descriptive Statistics**

- Min / Max, Mean, Median (approx.)
- Quantiles (approx.)
- Standard Deviation
- Variance
- Correlation
- Covariance
- Sum of Squares (cross product matrix for set variables)
- Pairwise Cross tabs
- Risk Ratio & Odds Ratio
- Cross-Tabulation of Data (standard tables & long form)
- Marginal Summaries of Cross Tabulations

### Statistical Tests

- Chi Square Test
- Kendall Rank Correlation
- Fisher's Exact Test
- Student's t-Test

### Sampling

- Subsample (observations & variables)
- Random Sampling

### **Predictive Models**

- Sum of Squares (cross product matrix for set variables)
- Multiple Linear Regression
- Generalized Linear Models (GLM) exponential family distributions: binomial, Gaussian, inverse Gaussian, Poisson, Tweedie. Standard link functions: cauchit, identity, log, logit, probit. User defined distributions & link functions.
- Covariance & Correlation Matrices
- Logistic Regression
- Classification & Regression Trees
- Predictions/scoring for models
- Residuals for all models

### Variable Selection

Stepwise Regression

### Simulation

- Simulation (e.g. Monte Carlo)
- Parallel Random Number Generation

### **Cluster Analysis**

K-Means

### Classification

- Decision Trees
- Decision Forests
- Gradient Boosted Decision Trees
- Naïve Bayes



- rxDataStep
- rxExec
- PEMA-R API Custom Algorithms



# SQL Server (in-database) R Services

# **SQL Server R Services**





### **In-database Analytics**

- R Services (in-database) Keep your analytics close to the data
- T-SQL Script Can be encapsulated in Stored Procedures
- Models are **built**, **trained**, **saved** as part of the ETL process (SSIS)
- Used for **batch** prediction (as part of the ETL process)
- Visual Studio SQL Database Project, Source Controlled, etc.
- Uses Microsoft ScaleR libraries



### Limitations

- Not supported in Azure SQL DB/DW, yet
- Not suitable for Interactive Data Science
- Only R, no python, yet.



# **SQL Server R Services**





### **T-SQL** Script

#### **Build and Save Model**

DECLARE @model varbinary(max);

EXEC sp\_execute\_external\_script @language = N'R'

```
-- Begin Learn Model Script
,@script =
N'
```

```
model <- lm(Output ~ Input, data = inputData);
print(summary(model))
modelbin <- serialize(model, NULL);</pre>
```

```
-- End Learn Model Script
```

, @input\_data\_1 =

```
Ν'
```

```
SELECT
```

```
Input,
```

Output FROM

```
FROM
```

```
demo.Data;
```

, @input\_data\_1\_name = N'inputData'

```
, @params = N'@modelbin varbinary(max) OUTPUT'
```

```
, @modelbin = @model OUTPUT;
```

```
INSERT INTO demo.Models (Name,Model,ModifiedDate)
SELECT 'regModel-demo-v2',@model,GETDATE()
```

#### Configure

Exec sp\_configure 'external scripts enabled', 1 Reconfigure with override;

### Model Summary

10 % - Messages

STDOUT message(s) from external script:

Call: lm(formula = Output ~ Input, data = inputData)

Residuals: Min 1Q Median 3Q Max -56.193 -13.545 -1.576 10.008 62.790

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 20.82 on 48 degrees of freedom Multiple R-squared: 0.8684, Adjusted R-squared: 0.8656 F-statistic: 316.6 on 1 and 48 DF, p-value: < 2.2e-16

### **Prediction Output**

	Results	Messages
	Input	Output
1	20	38.2391599966004

#### Prediction

DECLARE	@input_in FLOAT = 20;
DECLARE	<pre>@model_in VARBINARY(MAX);</pre>
SELECT	@model in = Model
FROM	demo.Models
HERE	ModifiedDate IN (Select MAX(ModifiedDate) FROM demo.Models);
EXEC SD	execute external script
@1	anguage = N'R'
c	5 5
Begin	n Predict
, as	cript =
۷.	
mod	<- unserialize(as.raw(model));
out	<pre>put &lt;- predict(mod, InputDataSet):</pre>
Inp	utDataSet\$output = output
data	a output = InputDataSet
pri	nt(data output)
End I	Predict
.@inp	ut data 1 = N'SELECT @Input AS Input:'
. Cout	put data 1 name = N'data output'
. @para	ams = N'@model varbinary(max).
. CI	@input float'
	.@model = @model in
	.@input = @input in
WITH	RESULT SETS (([Input] FLOAT, [Output] FLOAT)):



# Microsoft Analysis Services Data Mining

# **SQL Server Analysis Services**

# Data Mining

32

- Process data from many OLEDB and ODBC data sources
- Easy to build, interpret, deploy, and productionize
- SSIS Support Tasks to Train & Predict
- Interactive Visuals for model interpretation
- Excel Integration Data Mining Add-in

Limitations

- Limited Extensibility
- Limited Algorithms & Functionalities
- No Azure PaaS Service









- Sequence Clustering
- Time Series

# **SQL Server Analysis Services**



Microsoft

### **Visualizing Models**



# **SQL Server Analysis Services**



Server

Microsoft

### Excel Data Mining Add-in





D	17 -	E X V	fx						
4	A	8	с	D	E F	G	н	I	J
1				Key Influe	encers Report	for 'Purchase	ed Bike'		
2									
3	Key Inf	luencers and th	eir impact ov	er the values of 'Pur	chased Bike'	Discrimin	ation between f	actors leading	to 'No' and 'Yes'
4	Cile - La la la								
	Filter by Colum	nn' or 'Favors' to	see how vari	ous columns influen	nce 'Purchased Bike	Filter by 'Colur	mn' to see how o	different value	es favor 'No' or 'Yes'
5	Column ·	value	see how vari Favors	ous columns influen Relative Impact	ice 'Purchased Bike	Filter by Colum	nn' to see how o Value	Favors No 💌	es favor 'No' or 'Yes' Favors Yes
5 6	Column Colum	nn' or 'Favors' to Value 💌 2	see how vari Favors • No	ous columns influen Relative Impact	nce 'Purchased Bike	Filter by 'Colur Column Cars	mn' to see how o Value • 2	Favors No 💌	es favor 'No' or 'Yes' Favors Yes
5 6 7	Column  Cars Marital Status	nn' or 'Favors' to Value 2 Married	See how vari Favors V No No	ous columns influen Relative Impact	nce 'Purchased Bike	Filter by 'Colur Column Cars Cars	Value 2 0	Favors No 💌	es favor 'No' or 'Yes' Favors Yes
5 6 7 8	Cars Marital Status Region	nn' or 'Favors' to Value 2 Married North America	Ree how vari Favors V No No No	ous columns influen Relative Impact 🗗	nce 'Purchased Bike	Filter by 'Colur Column     Cars     Cars     Marital Status	nn' to see how of Value 2 0 Married	Favors No 💌	es favor 'No' or 'Yes' Favors Yes
5 6 7 8 9	Column Colum Cars Marital Status Region Cars	An' or 'Favors' to Value 2 Married North America	No No No No Yes	ous columns influen Relative Impact	ice 'Purchased Bike	Filter by 'Colur Column Cars Cars Marital Status Marital Status	Value 2 0 Married Single	Favors No 💌	es favor 'No' or 'Yes' Favors Yes 👻
5 6 7 8 9	Column Column Cars Marital Status Region Cars Marital Status	An' or 'Favors' to Value 2 Married North America 0 Single	No No No Yes Yes	ous columns influen Relative Impact	ice 'Purchased Bike	Filter by 'Colur Column Cars Cars Marital Status Marital Status Cars	Value 2 0 Married Single	different value Favors No 💌	es favor 'No' or 'Yes' Favors Yes 👻
5 6 7 8 9 10	Column Column Cars Marital Status Region Cars Marital Status Cars	Inn' or 'Favors' to Value 2 Married North America 0 Single 1	No No Yes Yes Yes	ous columns influer Relative Impact	ice 'Purchased Bike	Filter by 'Colum Column Cars Cars Cars Marital Status Cars Cars Region	Value 2 2 0 Married Single 1 Pacific	different value Favors No 💌	es favor 'No' or 'Yes' Favors Yes 👻



# **Azure Cognitive Services**

# Azure Cognitive Services



### HITACHI **Inspire the Next**

### Ready-to-use Intelligence

#### Language

Allow your apps to process natural language, evaluate sentiment and topics, and learn how to recognise what users want.



#### Language Understanding Intelligent Service

Teach your apps to understand commands from your users



#### Text Analytics API

Easily evaluate sentiment and topics to understand what users want



#### Web Language Model API

Use the power of predictive language models trained on webscale data



#### Bing Spell Check API

Detecting and correcting spelling mistakes in your app

### Speech

Processing spoken language in your applications



#### Bing Speech API

Convert speech to text and back again to understand user intent



#### Speaker Recognition API

Use speech to identify and authenticate individual speakers

#### Search

Make your apps, web pages and other experiences smarter and more engaging with the Bing Search APIs.



Q

#### **Bing Search APIs**

Search, image, video and news APIs for your apps

#### Bing Autosuggest API

Give your app intelligent autosuggest options for searches



Map complex information and data in order to solve tasks such as intelligent recommendations and semantic search.



0

#### **Recommendations API**

Predict and recommend items that your customers want





#### Face API

Detect, analyse, organise and tag faces in photos

by returning smart insights such as faces, images and emotion recognition.



•

#### Emotion API

Personalise user experiences with emotion recognition

Azure Cognitive Services



### HITACHI **Inspire the Next**

### Setup a Cognitive Services API

Cognitive Ser	vices APIs (preview)	□ ×	Create  Cognitive Services account - PREVIEW
Cognitive Services is curre Cognitive Services is a c that augment users' experi- With Cognitive Services y developed by experts in t	ently in preview. ollection of APIs that enable natural and contextual interaction with to riences via the power of machine learnt models from Microsoft. rou can tap into an ever-growing collection of powerful AI algorithms their fields-APIs including Academic Knowledge, Bing Autosuggest, Bi	pols	Account name     Enter a name for your API account     Subscription     Visual Studio Premium with MSDN
Search, Bing Speech, Bing Language Understanding Analytics, Translator Spee of Al-based tasks, giving	g Spell Check, Computer Vision, Content Moderator, Emotion, Face, I Intelligent Service(LUIS), Recommendations, Speaker Recognition, Te ech, Translator Text, and Web Language Model APIs. They simplify a vi you a quick way to extract insights from data.	xt ariety Academic	API type     Search by Api text.     Constant of the second
These APIs integrate into whatever language and platform you prefer. The APIs are o improving, learning, and getting smarter, so experiences are always up to date.		Bing Auto Bing Sear	osuggest API ch APIs
ש ל in ציי שייין שייי	8 🛛	Bing Spee Bing Spel	ch API
PUBLISHER	Microsoft	Compute Content N	r Vision API (preview) Moderator (preview)
USEFUL LINKS	More about Microsoft Cognitive Services Documentation Pricing Supplemental Terms of Use for Microsoft Azure Previews	Emotion / Face API / Language Recomme	peech service (Preview) API (preview) (preview) Punderstanding Intelligent Service (LUIS) (preview) endations API (preview) Recognition API (preview)
		Text Anal Translato Translato	ytics API (preview) r Speech API r Text API
Create		Web Lang	Juage Model API (preview) Create Automation options

#### https://www.microsoft.com/cognitive-services/



#### Put intelligence APIs to work

Microsoft Cognitive Services let you build apps with powerful algorithms using just a few lines of code. Thereads and platforms such as iOS, Android, and Windows, keep improving, and are

ଝ



# **Cognitive Features in Azure Data** Lake Analytics

# **Cognitive Features**

- Pre-built intelligence Text & Image Analysis
- Integrated with your data processing pipelines (DLA)
- Used for batch recognition (not singleton real-time)
- Scheduled & Automated using Azure Data Factory
- R & Python Extensions!
- Scalable Suitable for Big Data

### Limitations

- Limited Features
- Not suitable for real-time scoring



Data Processing & Patten









### First-time Installation

				Sam	ple Scripts		×	
				<b>⊼</b> Co	opy Sample Data	••• More		
H New Job     Resource group (change)     Kepocs-rq	xplorer 🛅 Delete 🖧 View All Jobs	Pricing tier Consumption		A	Sample Data Miss Click here to copy samp	<b>ing</b> ple data		
cunning contion ientral US		Default Data Lake Store kspocs Learn Explore sample scripts		8	Copying U-SQL extensi	ions files	Data Explorer 🛛 🖈 🗖 🗙	kspocs Data Lake Store
ubscription name ( <del>change)</del> risual Studio Premium with MSDN ubscription ID 30b5167-2934-4453-b323-8d28287ae375		Getting Started Explore interactive tutoria	ls		Query a TSV file			Filter New Folde
fanage View Job <i>Enter job url</i>	Go		Data Sources 1 도·		Create Database and Table	e	<ul> <li>Kspocs</li> <li>demo</li> <li>master</li> <li>Tables</li> </ul>	NAME
		Add User Wizard	kspocs (defau	$\bigotimes$	Populate Table		<ul> <li>Views</li> <li>Table Valued Functions</li> <li>Procedures</li> </ul>	🚞 contoso
				۵	Query Table		Assemblies      ExtPython	Cutputs
STATUS	JOB NAME	AUS	LANGU	AGE	DURATION		FaceSdk	sqlbits
Preparing	Install U-SQL Extensions - Reg	isterAll.usql 1 (0.4%	6) U-SQL		Just Now		imageEmotion ima ImageOcr	stream-data
							🖼 ImageTagging 🖼 TextCommon 🖼 TextKeyPhrase	🖿 usqlext

© Hitachi Consulting Corporation 2017. All rights reserved.

TextSentiment





# U-SQL Script

🕨 Submit Job 📄 Data Explorer 🛛 🗍 Oper	n File 🗕 Sav	ve As		
* Job Name 🖲	Priority 🖲	AUs 🛈		Estimated Cost <b>0</b>
dla-reviews-sentiment 🗸	1000		1	0.03 USD/minute
1 REFERENCE ASSEMBLY [TextComm	non];			
2 REFERENCE ASSEMBLY [TextSent	timent];			
3 REFERENCE ASSEMBLY [TextKeyF	phrase];			
4				
5 DECLARE @input_file string =	= "sqlbits/	input-data/book-reviews-sa	imple.c	sv";
6 DECLARE @output_file string	= sqibits	/output-results/reviews-se	entimen	t.tsv ;
/				
8 WINDUL_UALA =				
<sup>9</sup> EXTRACT Score decimal,				
11 EROM @input file				
12 USING Extractors (sv():				
13				
14 @sentiment =				
15 PROCESS @input data				
16 PRODUCE Score,				
17 Text,				
18 Sentiment string	ç,			
19 Conf double				
20 READONLY Score,				
21 Text				
22 USING new Cognition.Text	t.Sentiment	Analyzer(true);		
23				
24				
25 OUTPUT @sentiment				





## **Execution & Output**

dla-review	s-sentiment								
Resubmit	U Refresh	Duplicate Script	🛇 Cancel Job						
Job Summary Preparing	Queued	Running	Finalizing	Progress	~ ►	]	0 Os		
28s	18s	2min 5s	<b></b>				t it is very wel	I written. It covers a topic th Negative	-0.588958059409439
State	Succeeded					book-reviews-sample.csv	vorites; that's	not something I can say ab Positive	0.611546231721756
Duration	2min 51s						ault in our sta	ars is that our sun contains li Negative	-0.533393141312441
Author Submitted	the_flame_hea 3/23/2017, 2:	ad@hotmail.com 34:13 PM				SV1 Extract III 1 vertex → R: 3.12 MB	at defies its ge	enre in all the best ways pos Positive	0.560975730692288
Show more	re					(€) 1min 23s	anger among	the Nerdfighter community Negative	-0.557053135439017
Input Ou	utput					100%	asochistic thir	igs that you will ever do. Thi Negative	-0.5218289524489319
NAME						SV2 PodAggregate	begin with the	e author's assurance of bein Positive	0.53763207041451855
book-ravis	we-sample csv	3	12 MB			1 vertex → R: 3.27 MB     2s → W: 3.28 MB	is one. I want	ed very much to like it and f Negative	-0.537612193725128
DOOK-IEVIE	ews-sample.csv					5000 rows	/ for several m	nonths. I was vaguely aware Negative	-0.589140697376104
						$\downarrow$	her and YA lit	lover, so I was expecting gr Positive	0.51583341165808971
						reviews-sentiment.tsv	a negative re	view of this book as I am ve Negative	-0.639900366695452
							ad it 2.5 time	s since then. Every time I rea Positive	0.517728929394922
							ttention beca	use it felt like it was in the sa Positive	0.52928033676817154





## Scalable ML for Big Data

- Rich Spark ML Libraries
- Scalable, distributed, in-memory
- Extensible Python, R, Java, Scala
- Suitable for Big Data Batch Model Training and Scoring
- Spark Streaming for Real-time predictions
- Scheduled & Automated Using Azure Data Factory



HITACHI

**Inspire the Next** 

### Limitations

- Expensive to keep it up & running
- Slow to spin-up





# Spark ML Pipelines

Spark ML standardizes APIs for machine learning algorithms to make it easier to combine multiple task into a single pipeline, or workflow.



HITACHI

**Inspire the Next** 

- Transformers used for data pre-processing. Input: DataFrame Output:DataFrame
- Estimators ML algorithm used to build a predictive model. Input: DataFrame Output: Model.
- Parameters Configurations for Transformers and Estimators
- **Pipeline** Chains Transformers and Estimators





### HITACHI Inspire the Next

# Spark ML Functionality



#### Transformers

#### **Text Feature Extraction**

- TF-IDF (HashingTF and IDF)
- Word2Vec
- CountVectorizer
- Tokenizer
- StopWordsRemover
- n-gram

#### Feature Selection

- VectorSlicer
- RFormula
- ChiSqSelector

#### **Dimensionality Reduction**

PCA

#### **Features Vector Preparation**

- VectorAssembler
- VectorIndexer
- StringIndexer
- IndexToString

#### Feature Type Conversion

- Binarizer
- Discrete Cosine Transform (DCT)
- OneHotEncoder
- Bucketizer
- QuantileDiscretizer

#### Feature Scaling

- Normalizer
- StandardScaler
- MinMaxScaler

#### **Feature Construction**

- SQLTransformer
- ElementwiseProduct
- PolynomialExpansion

### Estimators (supervised)

#### Classification

- Decision Trees Ensembles
- Naïve-Bayes
- SVM

#### Regression

- Linear Regression
- SVM

### **Other (Unsupervised)**

- Clustering
- Collaborative Filtering
- Frequent Pattern Mining

Spark ML - Example

from pyspark import SparkContext
from pyspark.sql import SQLContext







<pre>from pyspark.sql.types import *</pre>	_
from pyspark.ml import Pipeline	
from pyspark.ml.classification import DecisionTreeClassifier	
from pyspark.ml.feature import StringIndexer, VectorIndexer	
<pre>from pyspark.ml.evaluation import MulticlassClassificationEvaluator</pre>	
# Set up spark and sgl contexts	
<pre>snarkContext = SnarkContext('spark'//headnodehost'7077' 'nyspark')</pre>	
salcontext = SOLCOntext(sparkContext)	
Sizeeneere - Sizeeneere (Sparkeeneere)	
<pre>data_path = "adls:///ml_data/sample_data.txt"</pre>	
# Load the data stored in adls as RDD.	
<pre>data_text = sc.textFile(data_path)</pre>	
# parse data	
<pre>data_parsed = data_text.map(lambda line: line.split(' ')).filter(lambda row: row[0] =! 'N</pre>	arr,)
<pre>data_frame = sqlContext.CreateDataFrame(data_parsed,[])</pre>	
# Index labels, adding metadata to the label column.	
<pre>labelIndexer = StringIndexer(inputCol="label", outputCol="indexedLabel").fit(data_frame)</pre>	
# Automatically identify categorical features, and index them.	
featureIndexer = VectorIndexer(inputCol="features", outputCol="indexedFeatures", maxCateg	ories=4).fit(data_frame)
	<pre>model path = "adls:///ml models/sample model.mlm"</pre>
# Split the data into training and test sets (30% held out for testing)	
(trainingData, testData) = data_frame.randomSplit([0.7, 0.3])	# Save Model.
	model.save(model path)
# Train a DecisionTree model.	_
<pre>dt = DecisionTreeClassifier(labelCol="indexedLabel", featuresCol="indexedFeatures")</pre>	# Load Model.
	<pre>model = DecisionTreeClassifier.load (model_path)</pre>
# Chain indexers and tree in a Pipeline	
<pre>pipeline = Pipeline(stages=[labelIndexer, featureIndexer, dt])</pre>	# Make predictions.
	<pre>predictions = model.transform(testData)</pre>
# Train model. This also runs the indexers.	
<pre>model = pipeline.fit(trainingData)</pre>	# Select (prediction, true label) and compute test error.
	evaluator = MulticlassClassificationEvaluator(labelCol="indexedLabel", predictionCol="prediction", metricName="accuracy")
# Summary only.	accuracy = evaluator.evaluate(predictions)
<pre>treeModel = model.stages[2]</pre>	
print(treeModel)	# Print Accuracy
-	<pre>print("Test Error = %g " % (1.0 - accuracy))</pre>





# BigDL – Intel's Distributed Deep Learning Library

https://azure.microsoft.com/en-us/blog/use-bigdl-on-hdinsight-spark-for-distributed-deep-learning/





# **Concluding Remarks**



Interactive Data Science Studio Azure ML	Extensibility <ul> <li>Spark on HDI</li> <li>Azure ML</li> <li>Microsoft R Server</li> </ul>	<ul><li>Built-in Features</li><li>Azure ML</li><li>Spark on HDI</li></ul>	<ul> <li>Rich Model Interpretability</li> <li>SSAS Data Mining</li> <li>Microsoft R Server</li> </ul>
<ul> <li>Pre-built Intelligence</li> <li>Azure Cognitive Services</li> <li>Azure Data Lake Analytics</li> </ul>	<ul> <li>ML Pipelining</li> <li>Spark on HDI</li> <li>Azure Data Lake Analytics</li> <li>SQL Server R Services</li> <li>Data Mining SSAS</li> </ul>	Integration with Operational Apps Azure ML Azure Cognitive Services Microsoft R Operationalization	<ul> <li>Scalability (Big Data)</li> <li>Microsoft R Server</li> <li>Spark on HDI</li> </ul>

# **My Background**

Applying Computational Intelligence in Data Mining

- Honorary Research Fellow, School of Computing, University of Kent.
- Ph.D. Computer Science, University of Kent, Canterbury, UK.
- 28+ published journal and conference papers in the fields of AI and ML

HOME	QUESTIONS	JOBS					
	Khalid M PhD in Compute Data Scientist & Hitachi Consulti	agdy Sala r Science // Engineer // ng UK - Business	ama al : Insights & A	12.76 nalytics /			Add a new A
Overview C	ontributions	Timeline	Info	Stats	Scores	Netw	rork
Maria	The local			Posted	her Manuaget	_	Projects
Search by publica Ant Colony Op Order neural N	tion title or keywork	i lied to the re	egularizat	ion proces	s of a High		Article - 10     Chapter - 2     Conference Paper - 18
Conference Pag	Dec 2016 -	2016 IEEE Syr	mposium S	eries on Cor	nputational I		Other Research
Ashraf M. Abdell	oar - Khalid Mago	ły Salama					Full-texts - 6
Upload full-tex	t Add to pro	aject Add re	sources	<b>~</b>		-	Answers
Data reduction	for classificat	tion with ant	colony al	aorithms		_	Followers - 24 Otations - 171
Article Sep 2	016 - Intelligent I	Data Analysis	country at	gonulling		_	Open Reviews
Khalid Marsh Sa		-					

https://www.researchgate.net/profile/Khalid\_Salama

n SlideShare Search		
Most Recent * @ Select all	A A A A A A A A A A A A A A A A A A A	Q. Search Uploads
ST. LINE BEST FILMS	ST. LEW MORE STRUCT	
fing the Data Lake	Diservices & DevOps	oh Analytics
Mudd M. Bakere Bakeres Inspire Histori Constitu	Madd M. Sakore Baar kees inagete Histold Consulter	
Building the Data Lake with Azure Da 1 month ago, 78 slides	Microservices, DevOps, and Continuo 2 months ago, 55 slides	Graph Analytics 5 months ago, 85 slides
Edit • Preview •	Edit • Preview •	Edit =
ind Machine Learning		Proved Ontrackaged by:
Machino learning with Spark	Shark with HDInsight	Enterprise Cloud Data

#### https://www.linkedin.com/in/khalid-salama-24403144/

Search GitHub	Pull requests
	Overview Repositorie
	Popular repositories
	csharp-repo This repository contains my play demos C#
Khalid Salama khalid-m-salama	4 contributions in the last
A data scientist & engineer who is passionate about solving real-world problems using technology	Apr May Jui
<ul> <li>Hitachi Consulting UK</li> <li>London, UK</li> <li>khalid.m.salama@gmail.com</li> </ul>	Wed Fri Learn how we count contributio

HITACHI

**Inspire the Next** 

#### https://github.com/khalid-m-salama/sqlbits-2017



# HITACHI Inspire the Next