Proposal for the creation of course content for a Cloud and Big Data Workshop

# Summary

The proposal lays out a set of Hands on Labs and demo scenarios which will accompany slides that a trainer will use to deliver a day “Cloud and Big Data Camp”. The target audience will be architect/developers and the first iteration of the course will focus solely on Microsoft big data and messaging tools and frameworks which include Windows Azure Sql Database, HDInsight, Event Hub, DocumentDb, Storage, Streaming Analytics and AzureML.

The purpose of the course is to give a grounding to attendees in how lambda architectures can be applied using platform services in Azure as well as how machine learning can be incorporated in both real-time and batch-based systems.

# Delivery Output

Elastacloud will be responsible for the delivery of all specified labs and demos and additional support software which will be fully documented in .md format and available through a private github repository ownership of which can be transferred on successful acceptance of the course material. Each HOL/DEMO or software will be available in separate folders and fully documented in sets of setup steps. Where possible we’ll work in tandem with the slide author to ensure that the relevant HOLs/DEMOs are in keeping with the material being presented.

# Low Level Task Breakdown

The following table shows the raw tasks, description and time estimate for the completion of the lab/demo.

|  |  |  |
| --- | --- | --- |
| Title | Description | Time (days) |
| Data stream simulator with Event Hubs | A simulator application written in C# which will take in a set of files from a well-known directory and specifically to do with the HOL or demo context will parse line by line and place on an Event Hub that the trainer is using to present with. The software will take a set of command line parameters and use this to parse each file and push the file content onto the Event Hub at specified intervals so that they can be timestamped over a period. | 2 |
| Script to create consumer groups for each attendee, push this back to blob storage using stream analytics as passthru | A script will be provided to enable the creation of enough consumer groups for all of the attendees to have a single one. The standard tier of the Event Hub currently supports up to 20 so this may be sufficient in many cases. A demo script to create a storage account and then use this as an output in a stream analytics job. The data in this case will be read from Stream Analytics which will do a passthru query and put the data into one or more block blobs for consumption by batch application processing through HDInsight. | 2 |
| HOL: Explore the portal | A description of the new portal and documenting the creation of all of the things that will be used throughout the class, describing compute, storage, stream analytics, DocumentDb and using resource groups to couple application concepts and boundaries. | 1 |
| HOL: Create an HDInsight cluster with Remote Desktop | Create an HDI cluster through the Azure portal and remote desktop into the cluster head node, write up as a HOL. | 1 |
| HOL: HDI Batch Analysis and PowerBI | Batch analysis with aggregation on IoT sensor data doing simple sums, counts and averages of data using HIVE. Display the data through a simply generated Power Dashboard on the web. | 2 |
| Demo: Azure Data Factory to operationalize batch process | Demo script to watch for a fileset, spin up an HDInsight cluster, run a set of HIVE scripts, insert the results into a Sql Azure table and spin down the cluster | 1.5 |
| HOL: Create an HDInsight HBase cluster with PowerShell | Use powershell to spin up an HDInsight cluster and add an additional powershell CmdLet to run a HIVE query | 1 |
| HOL: Create HBase tables, Insert and query tables | Create an HBASE cluster and use HIVE serde to push data into an HBASE table while doing a simple transform and then query the data back through HIVE | 1.5 |
| HOL: DocumentDb | Create a DocumentDb and using combinations of C# and the shell build a simple worker role to pull data from the Event Hub and push into DocumentDb using the SDK. Write a simple aggregated query to display the data back to an HTML table in a web page (using either ASP.NET/node.js or another framework). | 2 |
| Demo: Azure Streaming Analytics simple application | Build a simple Tumbling Window streaming application which will aggregate IoT sensor data and push the results into a Sql Azure database table | 1 |
| Demo: Storm feeding data to simple dashboard and HBase | Using the EventHub spout read data off of a consumer group and write a bolt to aggregate data and write to HBASE. Write a simple web ui using a D3 chart which will read and display data from HBASE. | 2.5 |
| HOL: Build a simple decision tree/forest in AzureML | Setup an AzureML namespace, import data, train dataset using a decision forest, replace with a decision jungle to show how to compare classifiers and see which one is better for predicting an aspect of the dataset which is being predicted. Expose s a web service and show a simple Request/Response usage. | 2.5 |

# High Level Task Breakdown

|  |  |
| --- | --- |
| Activity | Time |
| Dataset selection and preparation | 1 day |
| Total task time | 21 days |
| Contingency and iteration improvements | 6 days |
| Project management (@10%) | 2 days |
| TOTAL | **30 days** |

# Delivery Schedule

A suitable dataset will be agreed up front for all labs although this may entail the selection of several datasets. Datasets will either be created by Elastacloud or some cases taken and pre-prepared from the available open source datasets.

Elastacloud will delegate Richard Conway as the project manager and David Crossland (and Andy Cross) to deliver the course material.

|  |  |
| --- | --- |
| Activity | Date |
| Dataset selection and preparation | 14/1/2015 |
| Data Stream Simulator | 15/1/2015-16/1/2015 |
| Consumer Group and SA Scripting | 19/1/2015-20/1/2015 |
| HOL: Explore the portal | 21/1/2015 |
| HOL: Create an HDInsight cluster with Remote Desktop | 22/1/2015 |
| HOL: HDI Batch Analysis and PowerBI | 23/1/2015-26/1/2015 |
| Demo: Azure Data Factory to operationalize batch process | 27/1/2015-28/1/2015 |
| HOL: Create an HDInsight HBase cluster with PowerShell | 28/1/2015-29/1/2015 |
| HOL: Create HBase tables, Insert and query tables | 29/1/2015-30/1/2015 |
| Feedback and changes | 2/2/2015-3/2/2015 |
| HOL: DocumentDb | 4/2/2015-5/2/2015 |
| Demo: Azure Streaming Analytics simple application | 6/2/2015 |
| Demo: Storm feeding data to simple dashboard and HBase | 9/2/2015-11/2/2015 |
| HOL: Build a simple decision tree/forest in AzureML | 11/2/2015-13/2/2015 |
| Feedback cycle and completion | 16/2/2015-19/2/2015 |

End of day reports will be provided to stakeholders and optional scrums (although given the time difference not sure how effective this can be) to show progress and an hour weekly call showing progress, highlighting issues and agreeing scope and next steps where necessary.