**Final Project**

**Web Server Log – Clickpath Analytics**

**Problem Statement**:

Ecommerce retailers and other companies who have an online presence are trying gather more details about the customers browsing or online shopping patterns, the products they buy, the products they may be interested in the future and also provide a better shopping experience. They perform Basket Analysis, Path optimization and even try to analyze the next product to buy.To achieve this, companies have to process massive amounts of data sets in terms of web server logs which is also referred to as Clickpath or Clickstream data. This information is captured by Webserver as customer navigates around the website.

**Overview of technology**:

Azure HDInsight offers a cost-effective way to process massive amounts of data. Hadoops framework and its ecosystem helps to analyze this information easier, get better insights about the customer and help improve the effectiveness of the shopping. We use the tools and technologies provided to process large datasets of log files, get the required information from the logs and combine them with user profile and products data (these could be available from the OLTP application) to perform the required analytics. Hadoop offers multiple analytics tools for these big datasets. We will load, refine and visualize log data.

**High Level Steps:**

1. Created a HortonWorks Cluster available on Azure and configure it for SSH
2. Pre-process the Web logs sample data that is obtained
3. Create a SQL database in Azure
4. Move the sample data into Hadoop File system
5. Use Pig latin script to combine the web server logs into one.
6. Use Sqoop to move the data from traditional RDBMS (OLTP system) to Hive
7. Create custom Hive tables that will store the final data that is created for visualization.
8. Use Zepplin and Power BI for Analytics and visualization

**Data set obtained at**:

<https://s3.amazonaws.com/hw-sandbox/tutorial8/RefineDemoData.zip>

**Hardware/OS**:

Intel Core i5-5300U CPU 2.30 GHZ, 16 GB RAM, 64 bit Windows 7 operating system

**Software used**:

SQL Server Management studio, PowerBI, Powerview

**Two minute (short):** <https://youtu.be/eaO8d-c93G8> **| 15 minutes (long):** <https://youtu.be/IYv9PkPdCjk>