International Journal of Database Theory and Application
Vol.9, No.6 (2016), pp.191-198
http://dx.doi.org/10.14257/ijdta.2016.9.6.19

Analyze NYC Taxi Data Using Hive and Machine Learning

Bayan Alghuraybi[1], Krishna Marvaniya[2], Guojun xia[3] and Jongwook Woo[4]

[1,2,3] Grad Student, [2] Prof., Department of Computer Information Systems, California State University Los Angeles

[1] balghur@calstatela.edu, [2] kmarvan@calstatela.edu, [3] gxia@calstatela.edu, [4]jwoo5@exchange.calstatela.edu

Abstract

Machine learning utilizes algorithms to run predictive models that learn from a large dataset in an iterative manner. Predictive models are used in many business applications to gain competitive advantages and understand customers better. This paper concentrates on analyzing New York taxi trips and fares and presenting the methodology we used to address the problem and results reached by building through Azure Machine learning studio. Our practical approach starts with an exploratory analysis of NYC taxi data via Microsoft Power BI. Then more extensive analysis was conducted through Apache Hive data warehouse. Hive was built on top of Hadoop enabling data synopsis, query, and analysis. We implemented Hive queries to create tables in Microsoft Azure blob storage and store the data in external tables. Finally, we conducted our experiment by creating, training and testing the module. The finding and insights pertain to the main variables of our experiment: pick up time, drop off time and tip amount that could be integrated into an application and enhance business by picking the location with the highest tip for example.

***Keywords**: Machine learning, HDInsight, Hive, Cluster*

1. Introduction

The explosion of big data and predictive analytic solutions enable businesses to exploit patterns and identify opportunities more than ever. With predictive analytic, analyzing any available data is possible whether it was interactional data, attitudinal data, descriptive data or behavior data [1]. This was not possible few years ago without a help of data scientist and statistician that takes months. Now it becomes easier via solutions available on the cloud like Azure Machine Learning Studio. Not only prediction helps companies to acquire new customers but also retaining existing customers.

In our approach, we used a descriptive dataset: New York Taxi dataset that includes trip records from all trips finished in yellow and green taxis in 2013 (19G). This dataset includes pick-up and drop-off dates/times, pick-up and drop-off locations, trip distances, itemized fares, rate types, payment types and passenger counts. The dataset was collected and provided to the NYC Taxi and Limousine Commission (TLC) by technology providers authorized under the Taxicab &Livery Passenger Enhancement Programs (TPEP/LPEP) [2]. We used Hive and Azure Machine Learning, which is an easy and efficient way to analyze large datasets. First, we upload our datasets to HDInsight Hadoop cluster. Then, we use Hive to create database and tables. Finally, we used Reader module from Azure Machine Learning to proceed to model building and model deployment.

The rest of the paper is designed as follows: Section 2 covers Related work. Review of HDInsight, MapReduce, Hive and Azure Machine Learning Studio presented in Section 3. Ingesting data is explained in Section 4 followed by how data imported into Azure Machine Learning Studio in Section 5. Finally, we concluded in Section 6.

2. Similar Work

A lot of documentation and developers' forums assisted us to understand what data visualization tools and predictive method have been used recently. For example, Todd Schneider's used New York taxi data to make geographical maps with the help of PostgreSQL and Post GIS [3]. This experiment gave us a solid foundation how the data used for analysis by showing every taxi pickup and drop off in New York City from 2009 until 2015. This helped in conducting our study via Microsoft Azure. Jain, See and Shandilya have an online paper where they used initially scatter plot and 2-d histogram to aid the in representing the data via a heat map [4]. Because we used a completely different technology, more insight results from visualizing the data appear in our study. We used Microsoft Power BI to analyze, and visualize the New York taxi dataset for the first three months. Figure 1 shows the maximum amount of tips and passengers per location. As observed in Van Wyck Expy, the maximum tip paid was 17.70$ which is the highest and passengers numbers were the least comparing to the other location. Van Wyck Expy, Jamaica, NY 11430 is the address of John F. Kennedy International Airport (JFK) which explain why it has the highest tip amount.

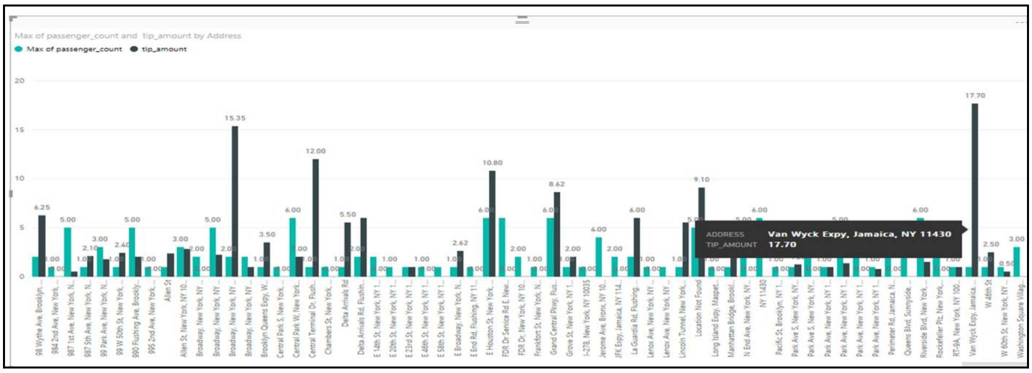


Figure 1. Maximum Tip Amount and Passengers Numbers with Respect to the Locations

3. Review of HDInsight, MapReduce, Hive and Azure Machine Learning Studio

3.1 HDInsight and Blob Storage

Hadoop is a distributed file system, this means storing large distributed chunks of data to different nodes. HDInsight is Hadoop on the azure cloud, a big data solution that basically deploying Hadoop on clusters with a monitoring mechanism aims to maintain availability, efficiency and ease of usage.

HDInsight consists of a head node and one or more data node with a replication storage capability that decrease the risk of losing the data. While Hadoop clusters are relying on traditional HDFS for storage, HDInsight and Blob storage maximize the HDFS efficiency by using HDFS as temporary storage for storing intermediate processed output and temporary job's task. The data and results are actually stored in the Blob Storage that attached to the HDInsight cluster.

Blob Storage is a storage deployed when HDInsight cluster created, allowing to drive the data to be stored in blobs and maximize the use of other Hadoop workflow. Using Blob Storage in Azure allows flexible scaling each storage efficiently. Hence, data stored on Blob Storage will not be discarded or lost even if the cluster is deleted. AzCopy is a Microsoft file copy utility that used to upload and transfer data to and from the Blob Storage via command line. After downloading the executable file and changing the path to the AzCopy directory, it will allow to upload and download files by specifying the source and destination in the command. If the specified folder/container in the destination didn't exist it will simply create one.

Beside Azure Blob storage, HDInsight contains dashboard in Azure portal. It provides console where developers can execute Hive queries, opening remote desktop connections and execute MapReduce jobs. Also, displays an overview of the cluster performance, usage and information in an illustrated graph. The console allows monitoring the running applications in a real time frame alarming for any abnormal activity. Thus, HDInsight is the optimal solution when it comes to data cleaning, normalizing and generating reports using advanced analytic tools and machine learning technologies [5].

3.2. MapReduce

In order to process data stored in HDFS, MapReduce is a programming model for splitting data into independent chunks that processed in parallel over a cluster of computers. MapReduce relies on two concepts: Mapper and Reducer. In Mapper, the desired information is extracted from the dataset, then the values are organized and mapped to unique keys. On the Reducer part, all the outputs of the mapper are sent to the reducer for processing each information to a specific key.

3.3. Hive

Hive is built as data warehouse solution above the Hadoop system with an aim to run and execute queries [5]. This framework is available as interactive HDInsight console or as Command Line in the Remote desktop RD to execute MapReduce jobs across the cluster through Hive SQL statements. While Hive is SQL-like queries that allow analyzing data easily, CLI at the head node in RD that saves execution time and possibly makes a difference for some businesses and applications.

3.4 Business Power BI

A platform of business analytic tools for visualizing, querying and building an informative report that brings more insights to the data being analyzed in a compelling way at any platform: tablet, phones, and computers. Enabling connecting, exploring, monitoring, sharing and synchronizing all the data that matter to the business across different data source. The service is available in the cloud or on premise maximizing analytic capability for massive scale and even streamed data.

Power BI can be integrated with SQL server database, analysis technology, and different applications. Power BI map is provided via an integration with Bing. Using specific Latitude and Longitude in decimal format to get to the right location. However, current it has a certain limitation when it comes to the precise coordination of longitude and latitude as Bing will view data on the state level, not the city level. A better geocoding approach can overcome this limitation by using Query Editor in Power BI Desktop [6].

3.5. Azure Machine Learning Studio

It is a Predictive analysis solution built based on the idea of running, editing, saving, experiment. It can be published as web services to be part of a web application. It requires some data science knowledge like statistic and programming. Until the trained model reaches the level of satisfaction and considers an effective model, an iterative process is used by modifying various functions and picking the suitable algorithms. In short, Azure Machine Learning Studio is all the understanding of Machine learning capability transforms as services to be consumed and developed. Replacing all thousand lines of code to be able to focus on the training model and to perform the necessary changing to reach the desired model [7].

4. Microsoft Azure Ingest data

4.1. Load Data Into Storage Environments for Analytics

After creating the HDInsight cluster and specifying the storage and nodes need for the experiment. We download New York City Taxi Dataset in a local device in order to upload it to the Blob Storage of the HDInsight [8]. The main goal from uploading the data to storage was to get the advantage of the data locality in Hadoop system. The data will be available even if we had to scale down the cluster later without the risk losing the data. To upload the New York City Taxi Dataset to the blob, the following command is applied in the AzCopy command-line [9]:

```
AzCopy/Source:<local_directory>/Dest:https://<account
name>.blob.core.windows.net<container_name> /DestKey:<account_key> /S
```

The source is the local directory where the dataset exists. The destination is the account associated with the blob storage. A destination key is required for accessing and downloading the data which is specified in the Blob storage setting. Note that /S is used to upload all the files including the sub files in the folder, means copy the whole directory structure.

4.2 Explore and Pre-Process Data through Business Power BI

Once the cluster is running and the dataset is uploaded were ready to the next step of analysis and query. However, Pre-processing and cleaning data first are important tasks that typically must be conducted before dataset can be used effectively for analysis in general and machine learning task specifically. Raw data is often noisy and unreliable, and may have missing values. So as a start we use PowerBI for exploring and analyzing data. This tool gives us more insight about drop-off and pickup locations as well as maximum tip amount per locations. Thus, it helps later in choosing the best algorithms that suit our NYC taxi data. One of the exploration we did with the dataset, we plot the total amount and the maximum number of passengers per location. As the figure above shows the maximum total amount paid. In this case, the above location got the maximum amount $134.90. As the Figure 2 below shows the maximum passengers and the total amount per locations. The total amount includes the fare and the tip for each trip. One thing to notice is 163 1st Avenue, New York, NY is only 2.7 miles away from Times square. Also, it is a business area that is full of hotels, dining restaurants and transit service available with an average waiting time of five minutes. Thus, it explains the high amount of fare comparing to the few number of passengers. Figure 3 is presenting the maximum total amount paid per location from during the day and compared with the evening shift. Obviously the trips during the day more than the night trips, but still some location where the number of the trips is equal whether it was night shift or not.

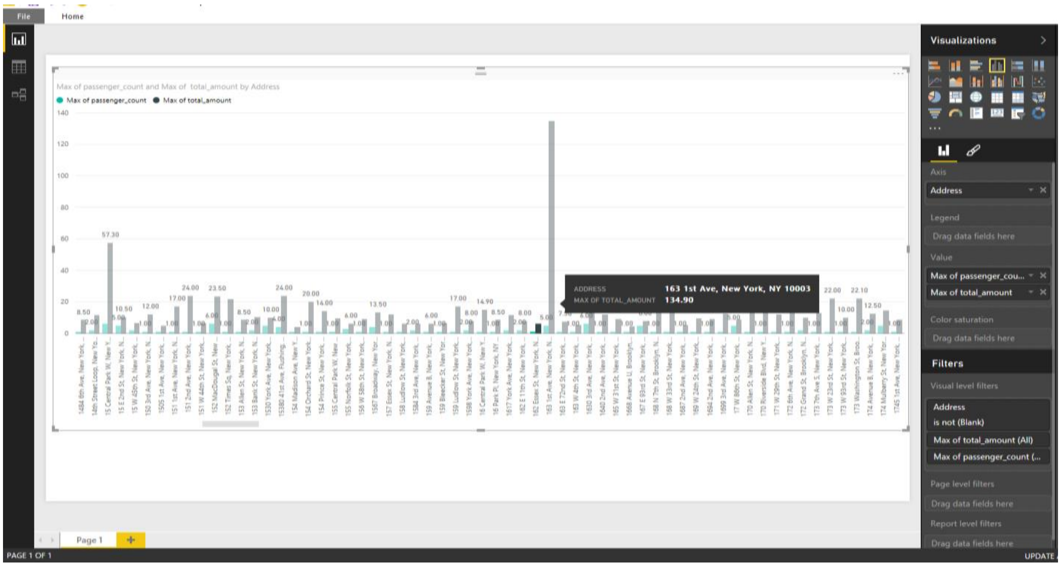


Figure 2. Maximum Passenger and Maximum Total Amount per Locations

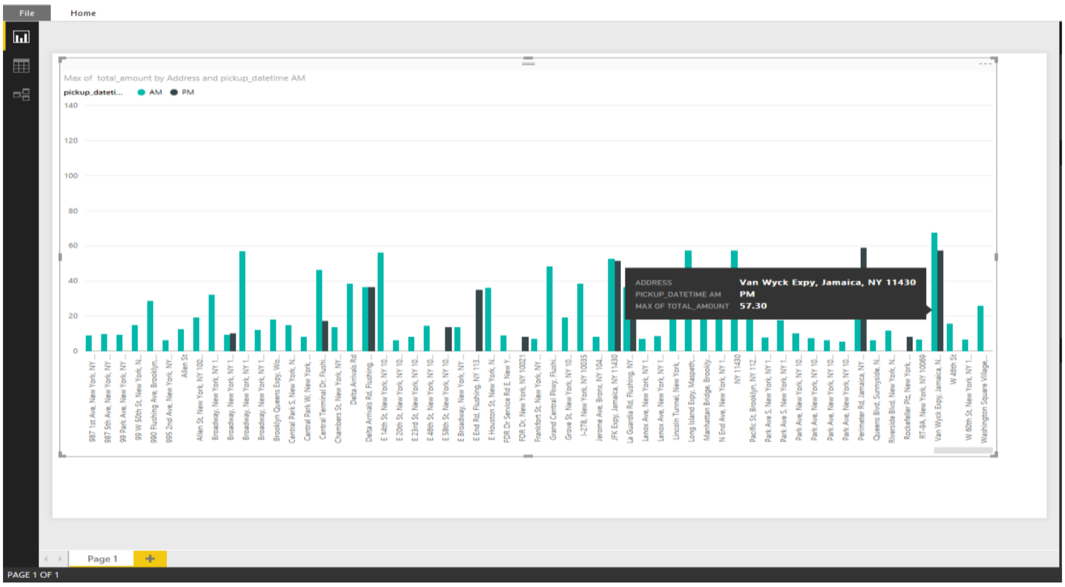

File
Home
Max of total_amount by Address and pickup_datetime AM
pickup_dateti... ● AM ● PM
140
120
100
80
60
40
20
0
ADDRESS Van Wyck Expy, Jamaica, NY 11430
PICKUP_DATETIME AM PM
MAX OF TOTAL_AMOUNT 57.30
Page 1
PAGE 1 OF 1

Figure 3. Maximum Total Amount per Locations Noon to Midnight

4.3 Explore and Pre-Process Data through Azure Hdinsight and Hive Queries

If the dataset you plan to analyze is too big, it is usually a good idea to down-sample the data to a smaller but representative to have more manageable size. This facilitates data understanding, exploration, and feature engineering. Its role in the Predictive Analytics Process is to enable fast prototyping of the data processing functions and machine learning models.

```
SELECT medallion, COUNT(*) as med_count
FROM nyctaxidb.fare
WHERE month<=3
GROUP BY medallion
HAVING med_count > 100
ORDER BY med_count desc
```

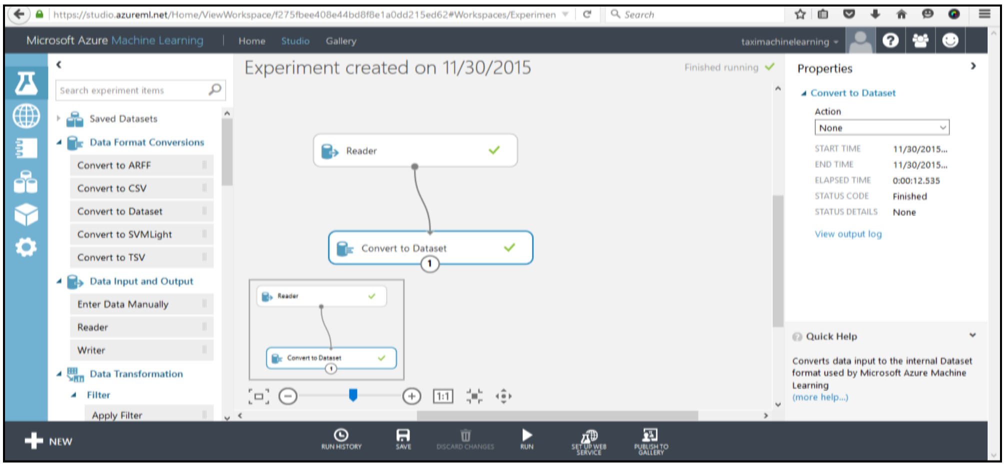
The code is illustrated in detail as follows:

1. Select medallion from the taxi fare table and count their occurrence.
2. Keep the condition as month should be greater than 3.
3. Group by the medallion and should have medallion count greater than 100.

The output should have medallion count from high to low.

5. Import Data into Azure Machine Learning Studio with the Reader Module

Figure 4 illustrates the workflow for conducting experiments in Azure Machine Learning Studio which consists of dragging-and-dropping components onto the canvas. Add the Reader module to the experiment, select the Data source, and then provide the parameters needed to access the data [10]. As the figure exemplifies on the left side under the Data Input and output section, we selected the reader module to ingest the data to the experiment and save it via writer module. What reader do is simply pointing to the Blob Storage that contains our data. Credential level and file format must be specified in the configuration of the reader. If any of the properties was wrong the reader will fail to start until all the required properties correctly specified. Then we convert the data for processing the missing values.


https://studio.azureml.net/Home/ViewWorkspace/f275fbee408e44bd8f8e1a0dd215ed62#Workspaces/Experimen

Experiment created on 11/30/2015

Finished running ✓

Properties

Convert to Dataset

Action: None

START TIME 11/30/2015...
END TIME 11/30/2015...
ELAPSED TIME 0:00:12.535
STATUS CODE Finished
STATUS DETAILS None

View output log

Quick Help

Converts data input to the internal Dataset format used by Microsoft Azure Machine Learning
(more help...)

Figure 4. Creating the Experiment Reader Module

5.1 Explore Data in the Predictive Analytics Process

```
SET sampleRate=<sample rate, 0-1>;
SELECT
   field1, field2, …, fieldN
FROM
   (
   SELECT
      field1, field2, …, fieldN, rand() as samplekey
   FROM <hive tsable name>
   )
where samplekey<='${hiveconf:sampleRate}
```

5.2 Create, Deploy & Consume Model

Azure Machine Learning enables to build, test, and deploy predictive analytics solutions [10]. Training module gives different output for classification and regression After refining the data, we split the data into modules to score the algorithm. We used Two-Class Logistic Regression to predict if the driver will get tipped or not. This algorithm used to predict one of the two states: tipped or not. For this algorithm, we need to exclude the other labeled fields like tip_class, tip_amount, and total_amount in the project module. Once the model has been trained with the existing data, it is ready for use to score new data. This model shows that AUC is 0.984 this indicates the classification model is fitted since AUC is fairly close to 1.

6. Conclusion

Azure Azure Machine Learning is a cloud-based predictive and analytics solution that can build models to be consumed as part of any web services. In our approach, we used Two-Class Logistic Regression algorithm to predict whether the driver will get tipped or not.

From the above experimental results, we can see:

- This model shows clients are more willing to tip.
- A Certain location tends to be more tipped than others *i.e.* Van Wyck Expressway New York, NY 11430 which is John F. Kennedy International Airport (JFK).
- People tend to be more likely to tip during the day compared to the night shifts.
- This model could take a further step to be used for web services so it helps taxi drivers to choose the best closest location where tip amount rate is the highest.

References

[1] IBM Predictive Customer Intelligence, http://www-01.ibm.com/software/analytics/media/smarter-paper/predictive-customer-intelligence/

[2] FOILing NYC’s Taxi Trip Data, http://chriswhong.com/open-data/foil_nyc_taxi.

[3] Analyzing-1.1Billion NYC-taxi-and-uber-trips-with-a-vengeance http://toddwschneider.com/posts/analyzing-1-1-billion-nyc-taxi-and-uber-trips-with-a-vengeance.

[4] J. Sahil, S. Alvin and S. Anish from University of California, San Diego, “Anticipating Taxi Tip-rates in NYC”, http ://jmcauley.ucsd.edu/cse 190/projects/sp15/050.pdf.

[5] Process Data with HIVE, http://hortonworks.com/hadoop-tutorial/how-to-process-data-with-apache-hive.

[6] How to use HDInsight to create cluster, https://azure.microsoft.com/en-us/documentation/articles/hdinsight-hadoop-tutorial-get-started-windows.

[7] What is Azure Machine Learning Studio?, https://azure.microsoft.com/en-us/documentation/articles/machine-learning-what-is-ml-studio.

[8] Import Data, https://msdn.microsoft.com/library/azure/4e1b0fe6-aded-4b3f-a36f-39b8862b9004.

[9] Transfer data with the AzCopy Command-Line Utility, https://azure.microsoft.com/en-us/documentation/articles/storage-use-azcopy/?cdn=disable.

[10] Deploy an Azure Machine Learning web service, https://azure.microsoft.com/en-us/documentation/articles/machine-learning-publish-a-machine-learning-web-service/#create-a-training-experiment.

[11] F. Nuno, J. Poco, H. T. Vo, J. Freire and C. T. Silva, “Visual exploration of big spatio-temporal urban data: A study of new york city taxi trips”, Visualization and Computer Graphics, IEEE Transactions on, vol. 19, no. 12, **(2013)**, pp. 2149-2158.

[12] R. Kankarej, P. Giri and J. Woo, “Data Analysis using Hive and Map-Reduce on Persons Obtaining Lawful Permanent Resident Status in USA”, in Journal of Science and Technology, vol. 5, no 11, **(2015)**.

[13] J. Mehta and J. Woo, “Big Data Analysis of Historical Stock Data Using Hive”, in Journal of Systems and Software, vol. 5, no. 2, **(2015)**, pp. 40-43.

[14] N. Bhattacharya and J. Woo, “Big Data Analysis of Airline Dataset using Hive”, in Journal of Systems and Software, vol. 5, no. 2, **(2015)**, pp. 22-26.

[15] J. Woo, “Market Basket Analysis using Spark”, in Journal of Science and Technology, vol. 5, no. 4, **(2015)**, pp. 207-209.

[16] Apache Hadoop, https://hadoop.apache.org.

[17] Connect Excel to Hadoop with the Microsoft Hive ODBC driver, https://azure.microsoft.com/en-us/documentation/articles/hdinsight-connect-excel-hive-odbc-driver.

Authors

Bayan Alghuraybi, is currently pursuing her Master's degree in Information Systems at California State University Los Angeles. She completed his Bachelor degree in Information Systems from King Abdul Aziz University, Saudi Arabia in 2012. Her interests include emerging technologies in Big Data and Analysis solution. Also learning about data processing, extracting patterns and information from them.

Krishna Marvaniya, is currently a student continuing her Master's degree in Information Systems at California State University, Los Angeles. She completed his Bachelor degree in Information Technology from Mumbai University, India in 2014. Her interests include programming, learning new technologies related to Big Data Analytics and web designing

Guojun Xia, is a graduate with a Master degree in Information Systems from California State University, Los Angeles. His interests include programming, Big Data Analytics tools and web development.

Jongwook Woo, is currently a Full Professor at Computer Information Systems at California State University Los Angeles. He is a director of the HiPIC (High-Performance Information Computing Center) at the university. He received his BS and the MS degree, both in Electronic Engineering from Yonsei University in 1989 and 1991, respectively. He obtained his second MS degree in Computer Science and received the PhD degree in Computer Engineering, both from University of Southern California in 1998 and 2001, respectively. His research interests are Information Retrieval /Integration /Sharing on Big Data, Map/Reduce, In-Memory Processing, and functional algorithm on Hadoop Parallel/Distributed/Cloud Computing, and n-Tier Architecture application in e-Business, smartphone, social networking and bioinformatics applications. He has published more than 40 peer reviewed conference and journal papers. He also has consulted many entertainment companies in Hollywood.