**Maz Saeed** Mobile: (952) 250-6004 Email: maz@datalogics.net

**SUMMARY**

I’m a skilled Cloud and Data Architect with 20 plus years in the industry. I have a successful track record of designing and implementing cutting-edge streaming and batch-oriented data pipelines on various cloud platforms – GCP, Azure and AWS – to support numerous business functions and use cases ranging from reporting to advanced analytics for companies in financial services, healthcare, retail, and manufacturing. With the data at many enterprises growing exponentially and stored in a variety of formats/disparate sources, my cloud and data architectures are designed to integrate data into one master platform that acts as a “single source of truth” by effectively harnessing power of on-cloud distributed cluster computing platforms like Kafka, Apache Spark, Airflow, Hadoop (Cloudera, Hortonworks), EMR or Elastic Search with primary objective of delivering key business intelligence insights faster and more accurately to facilitate data driven decision making at every level of the organization.

**KEY SKILLS**

* Design and implementation of SQL/NOSQL/dimensional data models for Business Intelligence and Advance Analytics applications optimized to run on a distributed computing cloud hosted and on-prem platforms – AWS, Cloudera
* Data lineage service development – source to target mapping of attributes and attribute transformation rules
* Data cataloging service development – implementation of automated metadata service to enforce file/attribute processing rules in terms of profiling data for valid values, ranges and referential integrity checks
* ETL service – coding ETL services in Python, Java and Shell scripts to cure harmonize and enrich data for downstream reporting and analytics consumption
* Extracting data out of un-structured documents like PDF, images with Optical Character Recognition (OCR) API such as Tesseract OCR and developing a Natural Language Processing (NLP) engine
* Expert in Docker image creation and deployment on a Kubernetes cluster
* Significant knowledge of building and deploying CI/CD pipelines with GitHub and Jenkins
* Expert in designing APIs in Java and Python deployed in a microservices architecture implemented with Kubernetes/Apache ZooKeeper orchestration layer
* Data Visualization and KPI dashboarding with Tableau, D3, R Shiny and PowerBI
* On-prem cloud -- Hortonworks/Cloudera multi-node cluster setup on RHEL/Debian Linux. Hadoop ecosystem Hive/SQOOP/OOZIE/IMPALA/Streaming, Apache Spark with Java/MLlib for Machine Learning application development
* Embedded Artificial Intelligence software development to optimize and automate decision making process -- includes dynamic in-memory experimental design and A/B testing
* Hosted cloud – AWS, S3, EC2, Lambda, Airflow, DynamoDB, Terraform, Serverless APIs
* Operating Systems: Debian/RHEL, Mac OS X, Windows Server
* Java J2EE Server – Multi-threading, collections API, connection pooling, web sockets

(JSR356), real-time Machine Learning algorithm development, REST based web service development

* Java build tools - ANT, Maven, SBT
* Version control systems – GitHub

**PROFESSIONAL EXPERIENCE**

**Hashmap Inc. 11/2020-Present**

**Principal Cloud Architect**

Responsible for architecting and cloud migration of both batch and streaming data pipelines for clients in the oil and gas industries. High level responsibilities include:

* Designing the reference architecture – batch and streaming
* Sizing of solution for current and future state of the application – estimate volume/velocity of data, ingestion/curation/enrichment, determine consumption patterns and support for concurrent usage
* Tech stack decisions – platform, storage backend and containerization options
* Designed data models to house dimensional and document structures (arbitrary attributes) – for curated and enriched data payloads
* Streaming/batch data pipeline workflow design and concurrent processing of workloads
* Coding, implementation, and testing of the proposed solution **–**
  + Coding of Kafka producer and consumer APIs to handle streaming workloads
  + Coding of various DAG operators in Python to trigger batch-oriented data pipelines

Tools/Technologies

* Streaming - Confluent Kafka multi-node cluster setup and configuration
* Azure Kubernetes Service (AKS) – Setup container registry, container deployment on AKS
* Batch – Airflow provision plus data pipeline DAG deployment in Airflow
* CI/CD pipelines – Setup and triggering of CI/CD pipelines in Gitlab and Azure Dev Ops (ADO)
* Snowflake –
  + Snowpipe development to ingest streaming
  + Change Data Capture (CDC) using Streams and Tasks in Snowflake
* Data Build Tools (DBT) pipeline with Snowflake as backend
* Lucidchart – Used in designing reference architecture and data modeling activities

**Dell Technologies – Sogeti Capgemini 6/2020-11/2020**

**Data Architect / Data Scientist**

Objectives: Re-architect a number of Decision Support tools both in commercial and consumer lines of businesses. Primary focus in this phase is to improve outcomes of commercial recommender model and consumer campaign response models in terms of more precise and timely delivery of “next best action” signals to key decision makers at various levels of the Dell enterprise

* Designed Apache Spark deployed as Microservices on Kubernetes cluster to process massive amount of data stored on Cloudera HDFS, S3 and other sources
* Optimized model training and deployment by pre-segmenting data into meaningful COHORTS and applying Machine Learning with grid search cost functions at the COHORT level
* Data models design optimized for HDFS deployment primarily consumed by Apache Spark jobs running on Kubernetes cluster
* Designed and implemented Apache Spark ETL pipelines to ingest data from sources like Teradata and Greenplum into HDFS and ECS/S3 file systems
* Designed and implemented data lineage and data cataloging services
* ETL pipelines deployed as DAGs triggered via Apache Airflow instance
* Designed and implemented Confluent Kafka streaming APIs – Producers and Consumers – as a Microservices architecture deployed in containers on Azure cloud

Tools/Technologies

* Interactive processing - PySpark running in client mode in Jupyter notebook
* Batch processing - Spark submit to Kubernetes cluster
* Logical and physical data model design with LucidChart software
* Microservices architecture design with LucidChart software
* Dev/Ops – build custom Docker images and deploy Docker containers on Kubernetes cluster
* Airflow Docker image creation and deployment in Kubernetes cluster
* CI/CD pipeline automation with GitHub
* Confluent Cloud managed services
* Kafka producers and consumers development in Java and Python

**United Airlines – IBM 2/2020-3/2020**

**Senior Cloud Data Architect/Developer**

Objectives and key responsibilities: Replace nearly a two decade old Revenue Management System with several known issues for calculating base fares to a more robust, real-time stream and event system deployed on AWS cloud. The new streaming Revenue Management System primarily designed to facilitate pricing and promotions decisions in real-time to maximize revenue potential for a given local/global market United operates in

* Designed logical models for raw/curated zones and deployed on AWS cloud optimized for DynamoDB and Redshift platforms
* Designed and implemented data lineage service responsible for mapping attributes from source to target data repository plus any business rules for transforming attributes
* Designed and implemented data cataloging service (per source) as configurable JSON object to enforce business rules for validating and processing attributes by the ETL service
* Coded ETL service to cure, enrich and harmonize data for downstream consumption
* Developed functional and technical microservices architecture and deployed microservices as decoupled streaming applications on a 3 node Kafka cluster to demonstrate the functionality of real-time base fare processing and calculations
* Created topics with partitions and replication factors for producer and consumer microservices
* Setup of on-prem producers to generate streams and in-cloud microservices to consume and process streaming data with Kafka Streams API

Tools/Techniques:

*Data Model* *design with LucidChart software:*

* Designed the microservices functional/technical architecture and detailed business process flows for each microservice
* Designed logical models and ERDs for raw and curated zones

**Three Bridge Consulting – Hagerty Insurance 10/2019-12/2019**

**Solution Architect/Data Architect**

Objectives and key responsibilities: With the premium quoting service riddled with inaccuracies due to data dispersed in several disparate systems and non-standardized metrics consumed as inputs to decision making, Hagerty faced sub-optimal premium quoting during underwriting process that led to revenue leakage. To alleviate data inconsistencies and address timely delivery of key insights for underwriting process, designed and deployed enterprise data hub/data lake on AWS platform by ingesting data from both structured and un-structured sources into hosted AWS environment and performing data cataloging, lineage, curation and enriching activities

* S3 bucket structure: Design S3 bucket structure for raw and curated landing zones and enforcing access controls on files via S3 tagging
* Data cataloging services: Design architecture in Lucid Chart and implement data cataloging services in Python to extract metadata during ingestion and store in a DynamoDB
* Data lineage services: Design architecture in Lucid Chart and implement data lineage services with D3 JavaScript library which included displaying complete data lifecycle from ingestion to curation and storing data in curated S3 buckets
* Data model design: Designed schema for curated data in S3 buckets for downstream reporting and machine learning consumption by various user groups

Techniques:

* Data ingestion pipeline: Airflow DAG, Lambda
* Data cataloging service: Python scripts running in an EC2 instance with DynamoDB backend trigged by Lambda
* Data Lineage: Python script development and execution triggered by Airflow DAG in EC2 to generate JSON for the D3 data lineage graph. D3 JavaScript, HTML, CSS, SVG development to display DAG on an EC2 hosted Python Flask web server
* Infrastructure provisioning: Terraform script to automate AWS environment deployment

**Mastech Digital – Highmark Healthcare 8/2019-9/2019**

**Data Architect/Data Scientist**

Objectives and key responsibilities: Deliver key insights extracted from un-structured data – images, PDF and other un-structured documents – via reporting and advanced analytics with the primary objective of enhancing stop-loss auto detection and premium quoting functions within the Highmark enterprise. High level steps:

* Convert PDF and Excel documents into images -- most documents had multiple pages
* Extract metadata and data out of images using character coordinates
* Convert data into well formatted JSON structure and store in HBase NoSQL table
* Perform searching and reporting functions on structured data stored in HBase table
* Predict probability of hitting stop-loss: Feature engineering, classifier model training/cross-validation and model deployment

Techniques:

* Designed the HBase data repository on a Cloudera CDH multi-node cluster as the backend for data pipeline APIs
* Us-structured to structured data:
  + Developed Python and Java code with Tesseract OCR (Optical Character Recognition) to extract text out of PDF and image files
  + Converted text into well-formed JSON structure and loaded into HBase NoSQL for downstream reporting and analytics functions
* Machine Learning: Turn un-structured text into feature space to feed into the classifier to compute probability of various claims submission events
* Display data metadata and lineage with D3 graph API

Tools:

* Tesseract OCR engine to extract text out of images running on multi-node Cloudera CDH cluster
* Python/Java to transform un-structured text into structured JSON objects
* HBase NoSQL database for structured JSON object data storage
* Serverless REST APIs to enable querying of data from a web client
* Stochastic Gradient Descent (SGD) classifier to predict probability with Apache Spark/MLlib
* Force Directed Graph development – coded Python scripts to generate JSON to act as a source powering the graph. Developed D3 JavaScript, HTML, CSS and SVG code to render the metadata and lineage as a Force Directed Graph accessed via a web client hosted on a web server

**Revolution Technologies – Cargill Corporation 12/2018-5/2019**

**Big Data Architect**

Objectives and key responsibilities:

* Developed plans for integrating data from four different ERPs and 1000+ plants scattered all over the world into a Hadoop Cloudera CDH multi-node cluster on deployed on AWS with the primary objective of delivering an enriched and harmonized view of data for reporting and analytics needs of the Cargill Animal Nutrition (CAN) enterprise
* Designed and deployed a dimensional data model on CDH data lake to present a harmonized view of the CAN business. Optimized the actual implementation based on anticipated data consumption patterns to fully leverage the distributed computing power of the cluster
* Led a team of data engineers in development of data pipeline work-flows that included following high-level data movement and enriching tasks:
  + Develop Kafka streams and SQOOP jobs to ingest data from ERPs into the base layer
  + Further cure and enrich data in staging layer with Spark jobs and Shell scripts
  + Transfer data into the harmonized dimensional model ready for business consumption
* Led on-going requirement status check discussions throughout the build with CAN business leads to ensure the final solution is in alignment with business expectations
* Developed User Acceptance Testing (UAT) execution plans

Techniques:

* Employed industry standard dimensional modeling techniques modified for scalable deployment optimized for handling volume and faster response times on a Cloudera CDH multi-node distributed computing environment deployed on AWS RHEL Linux distribution

Tools:

* ER/Studio Data Architect for designing the dimensional model
* Hadoop Impala for test plan and ad-hoc query execution
* Apache Spark with Python, Scala andShell scripts to automate various data pipeline ingestion and enriching tasks

**Solomon Page Group - Dollar General Corporation 2/2018 – 11/2018**

**Lead Data Scientist, Big Data Architect**

Objectives and key responsibilities:

* Define predictive analytics use cases – Propose techniques to implement weekly inventory demand forecasting model suite and recommender system development. Establish benchmarks for forecasting accuracy and expected lift from proposed modeling technique. Identify data sources, feature engineering and if use case is in POC (Proof of Concept) or roll-out to production stage
* Define architecture required for POC/role out – Design data architecture (data model) for local implementation or deployment on HDFS distributed multi-node node Cloudera Cluster
* Plan technology stack for Machine Learning – Activities include scoping size of data and memory requirements for the proposed algorithm – Python, Java versions running locally for POC or deployment stage with Apache Spark on a distributed HDFS Hadoop cluster
* Use case POC buy-in – Generally includes presenting findings of new technique for each use case in POC stage to stakeholders and get their approval before designing a production solution
* Design and deploy models in production – Python running locally is completely different from Python running on a distributed cluster with Apache Spark, significant effort and planning is required to convert POC into production code. Converted Python code to Apache Spark with Python, architected data repositories on HDFS to handle expanded scope of forecasting models

Techniques:

* Regression models – Support Vector Regression and Linear Regression
* Time-series models - ARIMA, Holt-Winters and Moving Averages
* Recommender Systems – Collaborative/Content Filtering and Hybrid models with Pearson Correlation, Cosine Similarity, Matrix Factorization

Tools:

* Python with Scikit-learn and Pandas libraries
* Cloudera implementation of Hadoop HDFS multi-node cluster on RHEL
* Hadoop ecosystem API stack – Yarn, Sqoop, Hive, NoSQL, Pig, MapReduce
* Apache Spark, MLlib Machine Learning with Java/Python
* Maven build tools for testing and deployment of Java and Python programs

**AIM Consulting - Best Buy 10/2016-11/2017**

**Data Scientist Consultant**

Objectives and key responsibilities:

* Design Measurement and Analytics framework for programs that fall under Strategic Growth Objectives (SGO) group
* Design demand forecasting models for new product launches with Monte-Carlo methods and other simulation-based optimization techniques
* Tasked with finding operational efficiencies in existing programs by employing Data Science and Machine Learning

Techniques:

* Developed Stochastic demand forecasting models using Monte Carlo Simulation to facilitate decision making for new product launches.  Incorporated sensitivity analysis into the forecasting process to determine key drivers of demand and the Net Present Value (NPV) of the expected revenue over the entire demand cycle.
* Developed Logistic Regression based Machine Learning models to predict returns and determine which model inputs best explain the returns phenomenon. The generated insights from this project will be used in fine-tuning sales force training to minimize returns.
* Formulated success criteria for pre and post program analysis with matched-pairs methodology to assess revenue lifts by comparing spend patterns of “similar” groups of customers in various time-periods.

Tools:

1. Base SAS plus SAS Enterprise Miner on Solaris/SunOS
2. Neteeza Data Warehouse
3. Unix shell scripts
4. VBA for simulation model prototype development
5. Java EE for server-side algorithm development

**AIM Consulting - Target Corporation 9/2015-7/2016**

**Big Data/Data Scientist Consultant**

Objectives:

* Leverage technology, employ pseudo automated quantitative methods with Machine Learning to optimize digital media spend and generate revenue lift through targeted online campaigns

Methodologies:

* ETL and distributed data storage: Designed and implemented a large scale (several Terabytes) data warehouse on Hortonworks HDFS multi-node cluster with full suite of Hadoop ecosystem – Yarn, MapReduce, Tez, Oozie, Sqoop and Hive
* Machine Learning: Primary objective is media spend and revenue optimization. In the process of developing/deploying self-calibrating models with Apache Spark – Scala and Java on Mahout/MLlib/Weka based Machine Learning

Tools: Hadoop HDFS multi-node cluster, Hive, Oozie, Apache Spark/Scala, Maven and SBT build tools

**Knowledgent Group, Inc. 10/2014–9/2015**

**Data Architect/Data Science Consulting**

**Zurich North America Insurance**

Objectives:

* For commercial accounts, develop ground up and residual pricing models to determine optimal premium based on account risks
* Develop claims propensity models to proactively identify accounts with high probability of a significant loss
* Assist actuaries with technical price development using data driven insights resulting from analytics

Methodologies:

* Modeled loss ratios with Weka decision tree (Reptree and M5P algorithms) machine learning systems

Tools: RapidMiner studio with Weka Machine Learning libraries and Radoop plug-in for RapidMiner for interacting with data on a Hadoop cluster

**WellCare**

Objectives:

* Which cohorts of members are at risk of dis-enrolling (aka churn) from the plans?
* Factors leading to dis-enrolment and turn them into actionable insights – the “why” part

Methodologies:

* Survival and Tree models in R using survival and rpart packages
* Data retrieval from Hadoop cluster with RMR based MapReduce jobs

Tools: Hadoop/HDFS (version: 2.5.0) on Linux x86\_64 (version 6.5) – 8 node cluster, Hue/Hive (version: 0.120-cdh5.1.2), R (version: 3.1.1) and R Studio (version: 0.98.1062) on 8 node Hadoop cluster

**Fidelity Investments**

Objectives:

* Optimized online user experience with data driven insights delivered through mining semi-structured text data generated by various channels – online, in-branch, phone, mobile etc.
* New customer acquisition – use analytics on semi-structured textual data to look for opportunities for improving conversion rates

Methodologies:

* Latent Dirichlet Allocation (LDA) and Support Vector Machines (SVM) based Machine Learning Topic model development in R/RHive/TM packages and Python with NumPy/Scikit-Learn
* Data retrieval from Hadoop cluster with Hive based MapReduce jobs
* Developed master shell script to automate on-going data processing and model calibration - staged on an edge-node of the Hadoop cluster

Tools: Hadoop/HDFS (version: 2.5.0) on Linux x86\_64 (version 6.5), Hive (version: 0.120-cdh5.1.2), R (version: 3.1.1) and R Studio (version: 0.98.1062), Python (version: 2.7.8) and Anaconda (version: 2.1.0)

**Aviana Global Technologies/IBM 4/2014–9/2014**

**Regional Client Director**

* Aviana, a partner of IBM, is a small consulting company specializing in Advanced Analytics consulting and Business Intelligence applications development
* Aviana at Boeing: Optimize the production operations for the F-15 fighter jet program
  + The primary objective is to identify risk factors in the supply/demand chain that would prevent on-time delivery of orders using SPSS Modeler

Methodologies:

* In a continuous test-learn-calibrate cycle, build various time-based states of the project plan, optimize and measure success rate
* Techniques used for model building include Regression, Decision Trees (C5.0 CHAID), and Neural Networks

**JP Morgan Chase Credit Risk – CRM Modeling and Analytics 9/2013–12/2013**

**Senior Risk Modeler CCAR**

Objectives:

* Develop and stress tested Risk models as part of the Comprehensive Capital Analysis and Review (CCAR) project in accordance with applicable Dodd-Frank standards
* Submitted recommendations to Model Validation and Management team, Federal Reserve and the OCC
  + The process involved assessing delinquency/default risk in various consumer mortgage portfolios stress tested for various scenarios to assist JPMC Portfolio Risk Management team in ascertaining true Risk Weighted valuation of assets and allocating cash reserves required to offset risk.

Methodologies:

* Calculated Probability of Default (PD) based on various test scenarios such as Housing Price Index (HPI) shifts over time and Regression analysis
* Computed Exposure At Default (EAD) and Loss Given Default (LGD) for each loan in a given portfolio
* Forecasted asset value based on Risk Weights computed from PD, EAD and LGD into the future quarters
* Conducted ELTV and Delinquency band QoQ migration analysis

Tools: R, Base SAS, SAS EG, Teradata, SQL Server, UNIX Shell Scripting, VBA

**Wells Fargo Home Mortgage 6/2012–7/2013**

**Analytics Consultant/Statistician**

Objectives: Use analytics to drive strategy and decision making

* Recommend optimal fee pricing strategies
* Improve fee accuracy on HUD statement, GFE to HUD variance analysis
* Establish admissible fee benchmarks by state/MSA

Methodologies:

* Developed Maximum Likelihood Estimators computed from Probability Mass Functions/Probability Density Functions to augment fee estimates
* Developed dynamic causal Regression based Linear models to communicate key dependent and independent variable relationships
* Designed and developed number of Data Warehouses (de-normalized) and ETL scripts with primary objective of mining data

Tools: Base SAS, SAS EG, Teradata, SQL Server, UNIX Shell Scripting, VBA

**IBM Advanced Analytics and Optimization 4/2010–5/2012**

**Lead Data Architect/Modeler**

Objectives: Optimize power flow on Hydro One grid with the goal of maximizing distributed power generation while keeping load within acceptable tolerance.

Methodologies:

* Designed and populated data repository in Oracle database with historic load, generation and weather data
* Developed Regression based load and generation forecasting models fitted to hourly load and generation curves
* Developed Mixed Integer Programming models to optimize power generation on the grid

Tools: SPSS Modeler, ILOG/CPLEX, Oracle PL/SQL

**Best Buy**

**Lead Architect/Modeler**

Objectives:

* Tier qualification: Evaluate and recommend Silver tier qualification for Best Buys’ RewardZone loyalty program by applying Action Cluster methodology. This exercise resulted in proposing an invite strategy based on insights deduced by evaluating a customer on a number of key attributes.
* Marketing mix: Produce a RewardZone targeted offer mailing list as part of a campaign to re-activate customer segments impacted by recent changes in the RZ program. Various customer cohorts were identified using Action Clusters who would likely respond to certain offers.
* Post Purchase Warranty Renewal Response Model: Who’s likely to respond to a warranty renewal offer? What’s their channel preference? How to time the offers to maximize response?

Methodologies:

* Feature Vector development and binning: Used as inputs to Action Clustering and developed from high priority business questions from the clients’ perspective
* Action Clusters scoring: Applied existing cluster scoring rules from a previously trained segmentation model to new customer data and produce profiling reports that include new feature vectors
* Developed BTP data warehouse, ETL scripts and renewal response model with LOGIT to predict probability of response to an offer
* Identified channel preferences by leveraging Action Cluster based segmentation. Timing of offer scenarios were constructed from historic solicitation data

Tools: SAS, Teradata, Oracle, SpeakEZ, UNIX/AIX Shell scripts, PERL, C++

**Information Resources, Inc. 8/2008-2/2010**

**Director, Customer Analytics**

Objectives: Implement Customer Centric Retailing for CPG retailers by identifying key customer purchasing behaviors in historic data evaluated on myriad of attributes such as products affinities, geographies, seasonality, ethnic skews etc. The insights were then transformed into customer strategy to assist a retailer with decisions such as:

* Store layout planning
* Localized assortment and placement of products in stores
* Enabling pricing and promotion decisions

Methodologies:

* Data preparation:
  + Performed ETL to create data repositories (One+ Terabytes) of sample data. ETL performed with Perl, Shell and compiled C scripts on UNIX/LINUX. Preferred DBMS - DB2 and Oracle in distributed data environment
* Data exploration:
  + Evaluated available variables in data and compute various descriptive statistics
  + Performed dimensionality reduction as necessary using techniques such as Market Basket and Principal Component Analysis
* Built and validated models: Identified homogenous customer segments (Trip Types) in data using K-Means Clustering, Decision Trees, Association Rules and Logistics Regression that are well classified and exhibit differentiable behavior. Train and validate models on sample data plus on-going model calibration and application to new data
* Preferred modeling tools were SAS and KXen

**Carlson Marketing Worldwide/Peppers & Rogers Group 2/2005-8/2008**

**Manager, Decision Sciences**

* Designed and executed CRM (Customer Relationship Management) initiatives for clients in financial services, retail and healthcare primarily to:
  + Increase customer retention and reduce churn
  + Optimize marketing budgets with targeted communications thereby reducing channel spend
  + Build customer loyalty and deliver higher ROI on various channel investments
* Data warehouse design and ETL (Extraction Transformation and Loading) script/batch job development for periodic data loads into data marts on SAS, Oracle and Teradata platforms
* Data Mining, Information delivery/Reporting and Measurement:
  + Supported Peppers and Rogers Group (acquired by Carlson) 1to1 CRM (Customer Relationship Management) initiatives such as the development and maintenance of the living Touchmap that describes various stages of customer lifecycles and interactions at different touchpoints (home grown web-based application developed using ASP/Dot Net)
  + Developed customer segmentation models (un-supervised K means)
  + Developed automated dashboards and reports of Key Performance Indicators (KPIs’), program measurement and ROI delivered via Flash enabled Web based dynamic dashboards developed on ASP and Dot Net platforms
* Improved the targeted Marketing Campaign’s response rates and program ROI
* Automated Business Rules notifications facilitated on-time and accurate delivery of information to various stakeholders

**KEY COURSES AND TRAINING**

* Hadoop + Data Scientist – Tech E Training Institute (April, ’14): Various components of this hands-on training include:
  + Configuring a Hadoop Cluster (HDFS) on CentOS/Linux plus loading files on HDFS
  + Developing Map/Reduce jobs in Java for processing large data files on HDFS
  + Implementing Machine Learning algorithms in Java with R and Mahout (Apache) libraries
* Machine Learning – Stanford University (Dec ’11): Supervised and un-supervised Machine Learning algorithms development and prototyping with Octave. Implementation of various cost functions e.g. Gradient Descent in Octave/R. Topics included: supervised learning (generative/discriminative learning, parametric/non-parametric learning, neural networks, support vector machines); unsupervised learning (clustering, dimensionality reduction, kernel methods); learning theory (bias/variance tradeoffs; VC theory; large margins); reinforcement learning and adaptive control.
* Java II Certificate Program – University of Minnesota (Nov '00): Building GUI based, multi-threaded Java Applets and Applications. Areas focused included JFC, Swing components, and OOP (Object Oriented Programming)
* Client/Server Technology Certificate Program - University of Minnesota (April '99); Scope of the program involved learning Client/Server architecture fundamentals, relational database analysis and design, and network communication systems to develop applications in a Client/Server environment using Visual Basic, PowerBuilder, SQL Server and Oracle.

**EDUCATION**

**University of South Dakota**

MBA/MIS in Econometrics and Information Systems

**University of South Dakota**

BS/BA in Finance and Economics