Ben Hodge

Data Scientist - Barclays Bank

New York, NY - Email me on Indeed: indeed.com/r/Ben-Hodge/7e0b49a303654fce

• 4+ years of experience in Data Scientist with strong technical expertise, business and leadership experience, and communication skills to drive high-impact business outcomes through data-driven innovations and decisions
• Extensive experience in Text Analytics, developing different Statistical Machine Learning, Data Mining solutions to various business problems and generating data visualizations using R, Python and Tableau
• Hands on experience on SparkMlib utilities such as classification, regression, clustering, collaborative filtering, dimensionality reductions
• Proficient in Statistical Modeling and Machine Learning techniques (Linear, Logistics, Decision Trees, Random Forest, SVM, K-Nearest Neighbors) in Forecasting/Predictive Analytics, Segmentation methodologies, Regression based models, Hypothesis testing, Factor analysis/ PCA, Ensembles
• Strong knowledge of statistical methods (regression, time series, hypothesis testing, randomized experiment), machine leaning, algorithms, data structures and data infrastructure
• Extensive hands on experience and high proficiency with structures, semi-structured and unstructured data, using a broad range of data science programming languages and big data tools including R, Python, Spark, SQL, Scikit Learn, R Shiny & Shiny Dashboards, Hadoop MapReduce
• Expertise in transforming business requirements into analytical models, designing algorithms, building models, developing datamining and reporting solutions that scales across massive volumes of structures and unstructured data
• Strong experience in Software Development Life Cycle(SDLC) including Requirements Analysis, Design Specification and Testing as per Cycle in both Waterfall and Agile methodologies
• Expertise in Technical proficiency in Designing, Data Modeling Online Applications, Solution Lead for Architecting Data Warehouse/Business Intelligence Applications
• Proficient in SQL, Database, Data Modeling, Data Warehousing, ETL and reporting tools
• Proficient in Data Science programming using Programing in R, Python and SQL
• Solid team player, team builder, and an excellent communicator

Willing to relocate: Anywhere
Authorized to work in the US for any employer

WORK EXPERIENCE

Data Scientist

Barclays Bank - New York, NY -

May 2016 to Present

Responsibilities

• Participated in all phases of data mining, data collection, data cleaning, developing models, validation, and visualization and performed Gap analysis
• Developed MapReduce/Spark Python modules for machine learning & predictive analytics in Hadoop Implemented a Python-based distributed random forest via Python streaming
• Performed Source System Analysis, database design, data modeling for the warehouse layer using MLDM concepts and package layer using Dimensional modeling

• Utilized domain knowledge and application portfolio knowledge to play a key role in defining the future state of large, business technology programs
• Created ecosystem models (e.g. conceptual, logical, physical, canonical) that are required for supporting services within the enterprise data architecture (conceptual data model for defining the major subject areas used, ecosystem logical model for defining standard business meaning for entities and fields, and an ecosystem canonical model for defining the standard messages and formats to be used in data integration services throughout the ecosystem)
• Used Pandas, Numpy, seaborn, SciPy, Matplotlib, Scikit-learn, NLTK in Python for developing various machine learning algorithms and utilized machine learning algorithms such as linear regression, multivariate regression, naive Bayes, Random Forests, K-means, & KNN for data analysis
• Provided the architectural leadership in shaping strategic, business technology projects, with an emphasis on application architecture
• Hands on database design, relational integrity constraints, OLAP, OLTP, Cubes and Normalization (3NF) and De-normalization of database
• Demonstrated experience in design and implementation of Statistical models, Predictive models, enterprise data model, metadata solution and data life cycle management in both RDBMS, Big Data environments
• Designed and implemented system architecture for Amazon EC2 based cloud-hosted solution for client
• Analyzed large data sets apply machine learning techniques and develop predictive models, statistical models and developing and enhancing statistical models by leveraging best-in-class modeling techniques
• Worked on customer segmentation using an unsupervised learning technique - clustering
• Utilized Spark, Scala, Hadoop, HBase, Kafka, Spark Streaming, MLlib, Python, a broad variety of machine learning methods including classifications, regressions, dimensionally reduction etc.
Environment: Erwin r9.6, Python, SQL, Oracle 12c, Netezza, SQL Server, Informatica, Java, SSRS, PL/SQL, T-SQL, Tableau, MLlib, regression, Cluster analysis, Scala NLP, Spark, Kafka, MongoDB, logistic regression, Hadoop, Hive, Teradata, random forest, OLAP, Azure, MariaDB, SAP CRM, HDFS, ODS, NLTK, SVM, JSON, Tableau, XML, Cassandra, MapReduce.

Data Scientist

Walmart - Minneapolis, MN -

October 2014 to May 2016

Responsibilities

• Designed and implemented email targeting feature in Walmart to send customers appropriate product recommendation
• Emails using collaborative filter technique.
Implemented algorithms to analyze credit card purchases to provide specialized recommendation to customers based on their purchase history
• Responsible for Interest clustering using SVM which clusters all the products that the user is interested in based on the search history and predictive window shopping or cart list
• Successfully designed and deployed instant online recommendations following hybrid approach which is derived by integrating CF and SPA (Sequential Pattern analysis)
• Worked with the team to improve spend analytics used in Walmart to categorize the items using Document clustering of NLP, also used lemmatization to avoid redundancy of same products in different categories
• Worked on Spark tool collaborating with ML libraries in eliminating a shotgun approach to understand customer buying patterns
• Used Hadoop platform to create market basket analysis to enable Walmart categorizing customers into groups or baskets, or products customers are more likely to purchase together

• Used Python language with the help of Scikit-learn libraries all the time for implementing the algorithms
• Worked with the team for Shopycat app to suggest users about buying ideal gifts to their friends during holiday season.
Environment: Snowflake, MySQL workbench, Hadoop HDFS, Mapreduce/YARN, HiveQL, Apache Sqoop, Apache ZOO, Apache, OOzie, Rstudio, Python, Theano, SQL, MS Excel 2016, Tableau, WINDOWS/Linux platform.

Big Data Analyst

Fairleigh Dickinson University -

January 2014 to September 2014

Responsibilities

• Created Hive external tables and designed data models in hive.
• Responsible for handling Hive queries using Spark SQL that integrates with Spark environment.
• Implementing YARN Resource pools to share resources of cluster for YARN jobs submitted by users.
• Data ingestion is done using Flume with source as Kafka Source & sink as HDFS.
• For one of the use case, used Spark Streaming with Kafka & HDFS/HBase to build a continuous ETL pipeline. This is used for real time analytics performed on the data.
• Performed import and export of large data set transfer between traditional databases and HDFS using Sqoop.
• Automated Sqoop Jobs in a timely manner for Data Migration from Existing RDBMS to HDFS using Shell Scripting.
• Worked with cloud services like Amazon Web Services (AWS) and involved in ETL, Data Integration and Migration.
• Performed transformations using Spark and then loaded data into HBase tables.
• Helping with performance tuning for Spark Steaming e.g. setting right Batch Interval time, correct level of Parallelism, selection of correct Serialization & memory tuning.
• Proactively monitored systems and services, manage backup and disaster recovery systems and procedure.
• Responsible for creating Hive tables, loading the structured data resulted from MapReduce jobs into the tables and writing Hive queries to further analyze the logs to identify issues and behavioral patterns.
• Implemented Hive Generic UDFs to handle business logic.
• Keep current with latest technologies to help automate tasks and implement tools and processes to manage the environment.
Environment: Hadoop, Map Reduce, HDFS, Spark, Hive, Java, Python, UNIX, Sqoop, HBase, Oracle, Cloudera Distribution, Oozie.

Hadoop Developer

Cerner HealthCare - Malvern, PA -

March 2013 to December 2013

Responsibilities

• Responsible for building scalable distributed data solutions using Hadoop
• Designed the projects using MVC architecture providing multiple views using the same model and thereby providing efficient modularity and scalability
• This project will download the data that was generated by sensors from the Patients body activities, the data will be collected in to the HDFS system online aggregators by Kafka

• Kafka consumer will get the data from different learning systems of the patients
• Spark Streaming collects this data from Kafka in near-real-time and performs necessary transformations and aggregation on the fly to build the common learner data model and persists the data in NoSQL store (HBase)
• Used Hadoop's Pig, Hive and Map Reduce for analyzing the Health insurance data to help by extracting data sets for meaningful information such as medicines, diseases, symptoms, opinions, geographic region detail etc.
• Developed workflow in Oozie to orchestrate a series of Pig scripts to cleanse data, such as removing personal information or merging many small files into a handful of very large, compressed files using pig pipelines in the data preparation stage
• Uses Pig in three distinct workloads like pipelines, iterative processing and research
• Uses Pig UDF's in Python, Java code and uses sampling of large data sets
• Involved in moving all log files generated from various sources to HDFS for further processing through Flume and process the files by using some piggybank
• Extensively used PIG to communicate with Hive using HCatalog and HBASE using Handlers
• Created Hive tables to store the processed results in a tabular format
• Good experience in PIG Latin scripting and Sqoop Scripting
• Involved in transforming data from legacy tables to HDFS, and HBASE tables using Sqoop
• Implemented exception tracking logic using Pig scripts
• Implemented test scripts to support test driven development and continuous integration
• Exported the analyzed data to the relational databases using Sqoop for visualization and to generate reports for the BI team
• Analyzed large amounts of data sets to determine optimal way to aggregate and report on it
• Good understanding of ETL tools and how they can be applied in a Big Data environment

Environment: Hadoop, Map Reduce, Spark, Kafka, HDFS, Hive, Pig, Oozie, Core Java, Python, Eclipse, HBase, Flume, Cloudera, Oracle, UNIX Shell Scripting.

EDUCATION

Bachelors in Computer Science in Computer Science

Cornell University

SKILLS

APACHE HADOOP OOZIE (3 years), APACHE HADOOP SQOOP (3 years), OOzie (3 years), Python (4 years), SQL (3 years)

LINKS

http://www.linkedin.com/in/ben-hodge-1073b2151

ADDITIONAL INFORMATION

TECHNICAL SKILLS

Machine Learning: Classification, regression, feature engineering, clustering, neural nets
Statistical Methods: Time Series, regression models, splines, confidence intervals, principal component analysis and dimensionality Reduction, bootstrapping

Programming languages: Python (panda, numpy, Scikit-learn), R, SQL, ML, Power BI, Excel
Selected Coursework: Machine Learning, Linear Algebra, Simulations, Probability and Statistics, Visualization, Big Data Analysis
Bigdata Ecosystems: Maprdeuce V2, HBase, HIVE, Sqoop, Oozie, Kafka
Spark Components: Spark Core 1.6, SparkSQL, Spark Streaming