**Avinash**

**Data Scientist**

**Email**: **avi.datascientist@gmail.com**

**Phone:** **469-453-6340**

**PROFESSIONAL SUMMARY:**

* Over 8 year's experience in Data Analysis, Data Profiling, Data Integration, Migration, Data Governance and **Metadata Management, Master Data Managementand Configuration Management.**
* Experience in various phases of Software Development life cycle (Analysis, Requirements gathering, Designing) with expertise in documenting various requirementspecifications, functionalspecifications, Test Plans, **Data Validation, Source to Target mappings, SQL Joins, Data cleansing.**
* Extensive experience in Text Analytics, developing different Statistical Machine Learning, Data Mining solutions to various business problems and generating data visualizations using **R, Python**.
* Expertise in transforming business requirements into analytical models, designing algorithms, building models, developing data mining and reporting solutions that scale across amassive volume of **structured** and **unstructured data.**
* Documenting new data to help source to target mapping. Also updating the documentation for existing data assisting with data profiling to maintain **data sanitation, validation.**
* Experience in conducting Joint Application Development (**JAD**) sessions for requirements gathering, analysis, design and Rapid Application Development (**RAD**) sessions Experience working on Data quality tools Informatica IDQ (9.1), Informatica MDM (9.1).
* Collaborated with the lead Data Architect to model the Data warehouse in accordance with FSLDM subject areas, **3NF format**, and **Snowflake schema.**
* Proficient in **SAS/BASE, SAS EG, SAS/SQL, SAS MACRO, SAS/ACCES.**
* Experience in end-to-end implementation of data warehouse project based on the **SAS EG.**
* Experience in extract data from adatabase such as DB2, Oracle, and **SME-IM, MAD, M240** and UNIX server using **SAS.**
* Extensive knowledge of Hadoop eco-system technologies like **Apache Pig, Apache Hive, Apache Sqoop, Storm, Kafka, Elastic Search, Redis, Flume** and **Apache HBase.**
* Experienced in analyzing data using **HiveQL** and **Pig Latin** and custom **MapReduce** programs in Java.
* Experienced in writing **Pig UDF** and **Hive UDF** and **UDAFs** in the analysis of data.
* Extensive experience in **Hive, Sqoop, Flume, Hue** and **Oozie**.
* Created practical predictive models with machine learning methods including Supervised (linear regression, logistic regression, decision tree, random forests, SVM etc.) and Unsupervised
* Knowledge in Business Intelligence tools like **Business Objects, Cognos, Tableau,**and **OBIEE**
* Experience with Teradata and big data as the target for data marts worked with **BTEQ, Fast Load** and **Multi-Load.**
* **Good knowledge in using all complex data types in Pig and MapReduce for handling the data and formatting it as required.**
* Integration Architect & Data Scientist experience in **Analytics, BigData, BPM, SOA, ETL** and **Cloud technologies.**
* Built Coe competencies in the area of **Analytics, SOA/EAI, ETL** and **BPM.**
* Experience in foundational machine learning models and concepts: **regression, random forest, boosting, GBM, NNs, HMMs, CRFs, MRFs, deep learning.**
* Experience in machine learning techniques and algorithms, such as **k-NN, Naive Bayes, SVM, Decision Forests, etc.**
* Good knowledge of Hadoop architecture and its components like **HDFS, MapReduce, Job Tracker, Task Tracker, Name Node and Data Node.**
* Collaborated with the lead Data Architect to model the Data warehouse in accordance with FSLDM subject areas, **3NF format, Snowflake schema**.
* Working knowledge of **DICOM** and Problem Loan Management applications.
* Highly skilled in using **Hadoop** (pig and Hive) for basic analysis and extraction of data in the infrastructure to provide datasummarization.
* Highly skilled in using visualization tools like **Tableau, ggplot2,**and **d3.js** for creating dashboards.
* Skilled in System Analysis, E-R/Dimensional Data Modeling, DatabaseDesign and implementing RDBMS specific features.
* Worked with HBase which is a **NoSQL**, column-oriented database.
* Experienced in importing and exporting data from therelational database into **HDFS** using **Sqoop.**
* Mapping and tracing data from system to system in order to establish data hierarchy and lineage.
* Using **Data Lineage** and reverse engineering as a way to track back errors in data till the data source.
* Experience in designing **starschema, Snowflakeschema** for **Data Warehouse, ODSarchitecture**.
* Experience in designing stunning visualizations using **Tableau** software and publishing and presenting dashboards, Storyline on web and desktop platforms.
* Having good experience in **NLP** with **Apache, Hadoop** and **Python**.
* Experience working with data modeling tools like **Erwin, PowerDesigner** and **ERStudio.**

## EDUCATION:

Bachelor of Computer Science

**TOOLS AND TECHNOLOGIES:**

|  |  |
| --- | --- |
| **Languages** | T-SQL, PL/SQL, SQL, C, C++, XML, HTML, DHTML, HTTP, MATLAB, DAX, Python |
| **Databases** | SQL Server 2014/2012/2008/2005/2000, MS-Access, Oracle 12c/11g/10g/9i and Teradata, big data, Hadoop |
| **Bigdata Ecosystem** | HDFS, PIG, MapReduce, HIVE, SQOOP, FLUME, HBase, Storm, Kafka, Elastic Search, Redis, Flume, Storm, Kafka, Elastic Search, Redis, Flume and Scoop. |
| **Statistical Methods** | Time Series, regression models, splines, confidence intervals, principal component analysis and Dimensionality Reduction, bootstrapping |
| **BI Tools** | Rsoft Power BI, Tableau, SSIS, SSRS, SSAS, Business Intelligence Development Studio (BIDS), Visual Studio, Crystal Reports, Informatica 6.1. |
| **Database Design Tools and Data Modeling** | MS Visio, ERWIN 4.5/4.0, Star Schema/Snowflake Schema modeling, Fact & Dimensions tables, physical & logical data modeling, Normalization and De-normalization techniques, Kimball &Inmon Methodologies |
| **Big Data / Grid Technologies** | Cassandra, Coherence, Mongo DB, Zookeeper, Titan, Elasticsearch, Storm, Kafka, Hadoop |
| **Tools and Utilities** | SQL Server Management Studio, SQL Server Enterprise Manager, SQL Server Profiler, Import & Export Wizard, Visual Studio.Net, Microsoft Management Console, Visual Source Safe 6.0, DTS, Crystal Reports, Power Pivot, ProClarity, Microsoft Office, Excel Power Pivot, Excel Data Explorer, Tableau, JIRA,SparkMLlib. |

**PROFESSIONAL EXPERIENCE:**

**Client: Proquest - MountainView, CA July 2017 – Till date**

**Role: Data Scientist**

**Description:** ProQuest LLC is an Ann Arbor, Michigan-based global information-content and technology company founded in 1938 as University Microfilms by Eugene B. Power. ProQuest provides solutions, applications, and products for libraries.

**Responsibilities:**

* As an Architect design conceptual, logical and physical models using Erwin and build datamarts using hybrid Inmon and Kimball DW methodologies.
* Worked closely with **business, datagovernance, SMEs** and vendors to define data requirements.
* Worked with data investigation, discovery and mapping tools to scan every single data record from many sources.
* Designed the prototype of the **Data mart** and documented possible outcome from it for end-user.
* Involved in business process modeling using **UML**
* Developed and maintained data dictionary to create metadata reports for technical and business purpose.
* Involved on Prediction model building, Machine Learning, Business process improvements, Visualization & Process implementation with R Programming and DeepSee
* Redesigned and developed SAS Applications with **Netezza** Database to the **Netezza** Applications reducing run time of Applications from 40 hours to 20 sec using PostgreSQL, nzsql, Aginity Workbench, SAS
* Created SQLtables with referential integrity and developed queries using **SQL, SQL\*PLUS, and PL/SQL.**
* Formulated procedures for integration of **R** programming plans with data sources and delivery systems and R language was used for prediction.
* Implementing SparkMlib utilities such as including classification, regression, clustering, collaborative filtering and dimensionality reduction.
* Design, coding, unit testing of **ETL** package source marts and subject marts using Informatica ETL processes for Oracledatabase.
* Developed Statistical Analysis and Response Modeling for Analytical Database contributors (logistic regression).
* Used **Pig** and **Hive** in the analysis of data.
* Used all complex data types in Pig for **handling data**.
* Created/modified **UDF** and **UDAFs** for Hive whenever necessary.
* Developed various **QlikView** Data Models by extracting and using the data from various sources files, **DB2, Excel, Flat Files** and **Bigdata.**
* Participated in all phases of datamining; datacollection, datacleaning, developingmodels, validation, visualization and performed Gapanalysis.
* data manipulation and Aggregation from adifferent source using **Nexus, Toad, BusinessObjects, PowerBI,**and **SmartView.**
* Handled importing data from various data sources, performed transformations using Hive, Map Reduce, and loaded data into HDFS.
* Interaction with Business Analyst, SMEs, and other Data Architects to understand Business needs and functionality for various project solutions
* Researched, evaluated, architected, and deployed new tools, frameworks, and patterns to built sustainable Big Data platforms for the clients
* Identifying and executing process improvements, hands-on in various technologies such as Oracle, Informatica, BusinessObjects.
* Loaded and transformed large sets of **structured, semi-structured** and **unstructured data.**
* Supported Map Reduce Programs those are running in the cluster.
* Managed and reviewed **Hadoop log files** to identify issues when ajob fails.
* Designed both **3NF** data models for **ODS, OLTP** systems and dimensional data models using Star and snowflake schemas.
* Developed **Pig UDFs** for preprocessing the data for analysis.
* Involved in **writing shell scripts**for scheduling and automation of tasks.

**Environment:** r9.0, Informatica 9.0, ODS, OLTP, Oracle 12c/11g, Hive, OLAP, DB2, Metadata, MS Excel, Mainframes MS Visio, Rational Rose, Requisite Pro., Hadoop, PL/SQL, SAS etc.

**Client: Agero – Boston, MA Apr 2016- Jun 2017**

**Role: Data Scientist.**

**Description:** We’re driving the next generation of vehicle and mobile technology forward, pushing the limits of big data to transform the entire driving experience. The majority of leading vehicle manufacturers and insurance providers use Agero’s innovations to strengthen their businesses and create stronger, lasting connections with their customers. Together, we’re making driving smarter and safer for everyone.

**Responsibilities:**

* Perform Data Profiling to learn about user behavior and merge data from multiple data sources.
* Implemented big data processing applications to collect, clean and normalization large volumes of open data using **Hadoop**ecosystems such as **PIG, HIVE,** and **HBase**.
* Designing and developing various machine learning frameworks using Python, R, and Matlab.
* Integrate R into Micro Strategy to expose metrics determined by more sophisticated and detailed models than natively available in the tool.
* Worked on different data formats such as **JSON, XML** and performed machinelearningalgorithms in Python.
* Worked as **Data Architects** and IT Architects to understand the movement of data and its storage and **ERStudio9.7**
* Processed huge datasets (over billion data points, over 1 TB of datasets) for data association pairing and provided insights into meaningful data association and trends
* Developed cross-validation pipelines for testing the accuracy of predictions
* Enhanced statistical models (linear mixed models) for predicting the best products for commercialization using Machine Learning **Linear regression** models, **KNN** and **K-means** clustering algorithms
* Participated in all phases of datamining,datacollection, datacleaning, developingmodels, validation, visualization and performed **Gapanalysis.**
* data manipulation and Aggregation from adifferent source using **Nexus, Toad, BusinessObjects, PowerBI,** and **SmartView.**
* Independently coded new programs and designed Tables to load and test the program effectively for the given POC's using with **Big Data/Hadoop**.
* Develop documents and dashboards of predictions in Microstrategy and present it to the business intelligence team.
* Developed various **QlikView**DataModels by extracting and using the data from various sources files, **DB2, Excel, Flat Files and Bigdata.**
* Good knowledge of HadoopArchitecture and various components such as HDFS, JobTracker, TaskTracker, NameNode, DataNode, SecondaryNameNode, and MapReduce concepts.
* As Architect delivered various complex **OLAP**databases/cubes, scorecards, dashboards and reports.
* Programmed a utility in **Python** that used multiple packages (**scipy, numpy, pandas**)
* Implemented Classification using supervised algorithms like LogisticRegression, Decisiontrees, KNN, NaiveBayes.
* Used **Teradata15** utilities such as **FastExport, MLOAD** for handling various tasks data migration/ETL from OLTP Source Systems to OLAP Target Systems
* Handled importing data from various data sources, performed transformations using Hive, MapReduce, and loadeddata into HDFS.
* Collaborate with data engineers to implement ETL process, write and optimized SQL queries to perform data extraction from Cloud and merging from Oracle 12c.
* Collect unstructured data from **MongoDB 3.3** and completed data aggregation.
* Perform data integrity checks, data cleaning, exploratory analysis and feature engineer using R 3.4.0.
* Conducted analysis of assessing customer consuming behaviors and discover thevalue of customers with RMF analysis; applied customer segmentation with clustering algorithms such as K-MeansClustering and Hierarchical Clustering.
* Work on outliers identification with box-plot, **K-means** clustering using **Pandas, NumPy.**
* Participate in features engineering such as feature intersection generating, feature normalize and Label encoding with Scikit-learn preprocessing.
* Use Python 3.0 (numPy, sciPy, pandas, sci-kit-learn, Seaborn, NLTK) and Spark 1.6 / 2.0 (PySpark, MLlib) to develop avariety of models and algorithms for analytic purposes.
* Analyze Data and Performed Data Preparation by applying thehistoricalmodel to the data set in AZUREML.
* Perform data visualization with **Tableau 10** and generate dashboards to present the findings.
* Determine customer satisfaction and help enhance customer experience using **NLP.**
* Work on Text Analytics, NaïveBayes, Sentiment analysis, creating word clouds and retrieving data from Twitter and other social networking platforms.
* Use Git 2.6 to apply version control. Tracked changes in files and coordinated work on the files among multiple team members.

**Environment:** ER Studio 9.7, Tableau 9.03, AWS, Teradata 15, MDM, GIT, Unix, Python 3.5.2, MLlib, SAS, regression, logistic regression, QlikView.

**Client: Capital One - Reston, VA. Dec 2014- Mar 2016**

**Role: Data Scientist.**

**Description:** Capital One Financial Corporation is a bank holding company specializing in credit cards, auto loans, banking and savings products headquartered in McLean, Virginia. Capital One is the eighth-largest commercial bank in the United States when ranked by assets and deposits and is ranked 9th on the list of largest banks in the United States by total assets.

**Responsibilities:**

* Data mining using state-of-the-art methods
* Extending company's data with third party sources of information when needed
* Enhancing data collection procedures to include information that is relevant for building analytic systems
* Processing, cleansing, and verifying the integrity of data used for analysis
* Doing ad-hoc analysis and presenting results in a clear manner
* Creating automated anomaly detection systems and constant tracking of its performance
* Strong command of data architecture and data modelling techniques.
* Hands on experience with commercial data mining tools such as Splunk, R, Map reduced, Yarn, Pig, Hive, Floop, Oozie, Scala, HBase, Master HDFS, Sqoop, Spark, Scala (Machine learning tool) or similar software required depending on seniority level in job field.
* Developed scalable machine learning solutions within a distributed computation framework (e.g. Hadoop, Spark, Storm etc.).
* Utilizing NLP applications such as topic models and sentiment analysis to identify trends and patterns within massive data sets.
* Knowledge in ML & Statistical libraries (e.g. Scikit-learn, Pandas).
* Having knowledge to build predict models to forecast risks for product launches and operations and help predict workflow and capacity requirements for TRMS operations
* Having experience with visualization technologies such as Tableau
* Draw inferences and conclusions, and create dashboards and visualizations of processed data, identify trends, anomalies
* Generation of TLFs and summary reports, etc. ensuring on-time quality delivery.
* Participated in client meetings, teleconferences and video conferences to keep track of project requirements, commitments made and the delivery thereof.
* Solved analytical problems, and effectively communicate methodologies and results
* Worked closely with internal stakeholders such as business teams, product managers, engineering teams, and partner teams.

**Environment:** Erwin r9.0, Informatica 9.0, ODS, OLTP, Oracle 10g, Hive, OLAP, DB2, Metadata, MS Excel, Mainframes MS Visio, Rational Rose, Requisite Pro. Hadoop, PL/SQL, etc..

**Client: Coventry Health Care, Minneapolis MN Apr 2012- Nov 2014**

**Role: Data Scientist.**

**Description:** Coventry Health Care Management utilizes multiple software systems to support the intake and processing of authorization requests, the exchange of data between the payer and vendors contracted to perform services on our behalf, manage Case and Disease programs, provide robust reporting and decision support, and generally automate and facilitate their business processes.

**Responsibilities:**

* Statistical Modelling with ML to bring Insights in Data under guidance of Principal Data Scientist
* Data modeling with Pig, Hive, Impala.
* Ingestion with Sqoop, Flume.
* Used SVN to commit the Changes into the main EMM application trunk.
* Understanding and implementation of text mining concepts, graph processing and semi structured and unstructured data processing.
* Worked with Ajax API calls to communicate with Hadoop through Impala Connection and SQL to render the required data through it .These API calls are similar to Microsoft Cognitive API calls.
* Good grip on Cloudera and HDP ecosystem components.
* Used ElasticSearch (Big Data) to retrieve data into application as required.
* Performed Map Reduce Programs those are running on the cluster.
* Developed multiple MapReduce jobs in java for data cleaning and preprocessing.
* Developed scalable machine learning solutions within a distributed computation framework (e.g. Hadoop, Spark, Storm etc.).
* Analyzed the partitioned and bucketed data and compute various metrics for reporting.
* Involved in loading data from RDBMS and web logs into HDFS using Sqoop and Flume.
* Worked on loading the data from MySQL to HBase where necessary using Sqoop.
* Developed Hive queries for Analysis across different banners.
* Extracted data from Twitter using Java and Twitter API. Parsed JSON formatted twitter data and uploaded to database.
* Launching Amazon EC2 Cloud Instances using Amazon Images (Linux/ Ubuntu) and Configuring launched instances with respect to specific applications.
* Exported the result set from Hive to MySQL using Sqoop after processing the data.
* Analyzed the data by performing Hive queries and running Pig scripts to study customer behavior.
* Have hands on experience working on Sequence files, AVRO, HAR file formats and compression.
* Used Hive to partition and bucket data.
* Experience in writing MapReduce programs with Java API to cleanse Structured and unstructured data.
* Wrote Pig Scripts to perform ETL procedures on the data in HDFS.
* Created HBase tables to store various data formats of data coming from different portfolios.
* Worked on improving performance of existing Pig and Hive Queries.

**Environment:** SQL/Server, Oracle 9i, MS-Office, Teradata, Informatica, ER Studio, XML, Business Objects, HDFS, Teradata 14.1, JSON, HADOOP (HDFS), MapReduce, PIG, Spark, R Studio, MAHOUT, JAVA, HIVE, AWS..

**Client: Spectra force Technologies India Pvt. Ltd - Hyderabad. Nov 2011- Mar 2012**

**Role: Data Architect/Data Modeler.**

**Description:** Spectra force is a leading global services firm that provides a portfolio of consulting, staffing and outsourcing services and solutions to a broad range of clients and industries worldwide. We are headquartered in Raleigh, NC, USA and have offshore global delivery centers in Pune, Chandigarh, and Hyderabad, India.

**Responsibilities:**

* Worked with large amounts of structured and unstructured data.
* Knowledge in Machine Learning concepts (Generalized Linear models, Regularization, Random Forest, Time Series models, etc.)
* Worked in Business Intelligence tools and visualization tools such as Business Objects, Tableau, ChartIO, etc.
* Deployed GUI pages by using JSP, JSTL, HTML, DHTML, XHTML, CSS, JavaScript, AJAX.
* Configured the project on WebSphere 6.1 application servers
* Implemented the online application by using Core Java, Jdbc, JSP, Servlets and EJB 1.1, Web Services, SOAP, WSDL.
* Handled end-to-end project from data discovery to model deployment.
* Monitoring the automated loading processes.
* Communicated with other Health Care info by using Web Services with the help of SOAP, WSDL JAX-RPC
* Used Singleton, factory design pattern, DAO Design Patterns based on the application requirements
* Used SAX and DOM parsers to parse the raw XML documents
* Used RAD as Development IDE for web applications.
* Preparing and executing Unit test cases
* Used Log4J logging framework to write Log messages with various levels.
* Involved in fixing bugs and minor enhancements for the front-end modules.
* Implemented Microsoft Visio and Rational Rose for designing the Use Case Diagrams, Class model, Sequence diagrams, and Activity diagrams for SDLC process of the application
* Doing functional and technical reviews
* Maintenance in the testing team for System testing/Integration/UAT.
* Guaranteeing quality in the deliverables.
* Conducted Design reviews and Technical reviews with other project stakeholders.
* Was a part of the complete life cycle of the project from the requirements to the production support
* Created test plan documents for all back-end database modules.

**Environment:** R 3.0, Erwin 9.5, Tableau 8.0, MDM, QlikView, MLLib, PL/SQL, HDFS, Teradata 14.1, JSON, HADOOP (HDFS), MapReduce, PIG, Spark, R Studio, MAHOUT, JAVA, HIVE, AWS.

**Client: Inspire Solutions - Hyderabad, Telangana Mar 2010- Oct 2011**

**Role: Data Analyst/Data Modeler.**

**Description:** Inspire is leading IT Company which considers ''IT Services'' as a way to develop business for the clients we work for. With over 150+ IT professionals from varied industries working across locations, we ensure nothing but best products and services with utmost care.

**Responsibilities:**

* Developed Internet traffic scoring platform for ad networks, advertisers and publishers (rule engine, site scoring, keyword scoring, lift measurement, linkage analysis).
* Responsible for defining the key identifiers for each mapping/interface.
* Clients include eBay, Click Forensics, Cars.com, Turn.com, Microsoft, and Looksmart.
* Designed the architecture for one of the first analytics 3.0. online platforms: all-purpose scoring, with on-demand, SaaS, API services. Currently under implementation.
* Web crawling and text mining techniques to score referral domains, generate keyword taxonomies, and assess commercial value of bid keywords.
* Developed new hybrid statistical and data mining technique known as hidden decision trees and hidden forests.
* Reverse engineering of keyword pricing algorithms in the context of pay-per-click arbitrage.
* Implementation of Metadata Repository, Maintaining Data Quality, Data Cleanup procedures, Transformations, Data Standards, Data Governance program, Scripts, Stored Procedures, triggers and execution of test plans
* Performed data quality in Talend Open Studio.
* Coordinated meetings with vendors to define requirements and system interaction agreement documentation between client and vendor system.
* Automated bidding for advertiser campaigns based either on keyword or category (run-of-site) bidding.
* Creation of multimillion bid keyword lists using extensive web crawling. Identification of metrics to measure the quality of each list (yield or coverage, volume, and keyword average financial value).
* Enterprise Metadata Library with any changes or updates.
* Document data quality and traceability documents for each source interface.
* Establish standards of procedures.
* Generate weekly and monthly asset inventory reports.

**Environment:** MS Office suite 2008, MS-SQL Server Management Studio 2000/2005/2008/R2, T-SQL, DTS, Replication, Rational Rose, Windows NT, MS SQL Reporting Services 2008, MS SQL Server Analysis Services 2008, MS SQL Server Integration Services 2008, MS Access, Erwin, SQL Query Analyzer.