**JAIDEEP SINGH**

Jaideep.singh2002@gmail.com

7837919090

Personal Statement

Highly versatile and results driven professional with 4.8 years of experience working within IT industry, including 2+ years as a Big data/Hadoop developer. Worked across several functions including service delivery and quality assurance. Experienced in providing solution design to clients, eliciting requirements and managing stakeholders.

**Professional Summary:**

* Total work experience is 5 years and relevant to Hadoop/ Big Data with 2 years of experience in Hadoop, Hadoop Ecosystem, HDFS, Data Warehousing and ETL tools
* Hadoop Developer with hands on experience on major components in Hadoop Ecosystem like Hadoop Map Reduce, HDFS, SPARK, HIVE, Hbase and Flume
* Efficient in building Python scripts using Python packages (i.e. PySpark, RDD, Data-frames, Data-sets, Numpy, Pandas, Request, JSON, date-time).
* Well versed in performing ETL from a wide variety of data sources using PySpark, Hive, Pandas and Numpy.
* Adept at designing, building and supporting pipelines of data ingestion, transformation, conversion and validation.
* Strong understanding of working with structured, semi-structured and unstructured datasets and Databases.
* An Agile mindset focused on continuous improvement and a team player.
* Proficiency in Exploratory Data Analysis, Data Wrangling and Visual representation for error checking.

Education

* Bachelors of Technology (Electronics & Communication)

**Employment History**

**Cognizant**

**Big Data Developer**  **Oct 2017 - Present**

As a Big Data developer my role includes building the connectivity to the data source. Create batch jobs to get data from different components in Hadoop environment. Translate functional requirements into technical design. Design and develop data ingestion programs. Monitor data ingestion processes end to end and optimize the overall data processing. Develop test scripts to validate data loaded in Hadoop platform.

Role and Responsibility:

* Extraction, transformation and loading using wide variety of data sources using PySpark, Hive, Scoop.
* Design, build and support pipelines of data ingestion, transformation, conversion and validation.
* Developed Python scripts for Data frames/SQL and RDD in Spark for Data Aggregation, queries and writing data back into RDBMS through Sqoop
* Created Hive schemas using performance techniques like partitioning and bucketing
* Use of custom UDF’s in Hive
* Enhancing the code using performance optimization techniques for Hive (optimized joins, JVM reuse etc) and PySpark (repartition/coalesce, compression )

**Data Analyst Aug 2016 - Oct 2017**

As a Data analyst my work involves acquiring raw data and getting it into a form ready for analysis. On day to day basis I was involved in activities including cleaning data, how to diagnose data for problems to dealing with missing values and outliers. Use Pandas to extract, filter, and transform data from DataFrames in order to drill into the data that really matters. Combine DataFrames to answer central questions.

**Key responsibilities:**

* Loading of data from different sources. Exploratory/Visual data analysis for errors or potential problems.
* Getting Tidy data, shaping it for analysis, combining datasets.
* Cleaning Data using Regex and creating functions for it. Testing data pro actively uses Assert statements.
* Import of different files types(flat file, Pickle, Spreadsheet, SAS/Stata, HDF5, MATLAB) with Custom import
* Use of SQLAlchemy for creation of Database engines, cursors, statement execution on different databases.
* Interacting with API's to get data from Web using urllib, requests and BeautifulSoup packages
* Data Exploration using Visual exploratory analysis via histograms, boxplots, scatter plots
* Reshaping and tidying data using techniques pivoting or melting and stacking or unstacking.
* knowledge of concepts such as upsampling, downsampling, and interpolation for time-Series data
* Data Merging/Concatenation, MultiIndexing, resampling