Subhajit Sen

Address : House No. 3, 10th A Cross, 16th Main,
BTM 2nd Stage, Bangalore - 560076
Email : subhajitsen220188@gmail.com
Mobile : 9880816126

ABSTRACT :

Pursuing my career as Data Scientist in implementing machine learning alongside big data stack solutions

LEAD DATA SCIENTIST:

Solution focused IT professional having 8+ years experience in solving complex data problems using Hadoop, Apache Spark and statistical analysis using machine learning algorithms and other ecosystem stacks. Data consumption, filtering, transformation, visualization, statistical modeling, analytical predictions and report generation workflow along with monitoring of distributed computing environment are the core responsibilities. Good understanding of complex processing needs of big data and has expertise in developing codes and modules to address those needs. Demonstrated capacity to identify root causes and direct lasting resolutions. Adept at articulating complex technical solutions to clients of varying technical understanding. Well versed with complex data structures, statistical models, algorithms, common operating systems and programming languages like Java, Scala and Python with a proven ability to master new tools and technologies.

AREAS OF EXPERTISE :

- Applied statistical analysis, machine and deep learning algorithms to build smart automation systems
- Developed stock advisory application using Linear regression, SVR and LSTM prediction models along with NLP to decipher social media and news feed sentiments
- Bond and equity data analysis using visualization and time series models
- Chabot implementation using NLP for department specific domain knowledge
- PDF, image and email document parsing and automating manual verification system using decision tree and other deep learning algorithms
- Existing legacy data architecture validation and proposal of new architecture with optimized data model along with big data stacks implementations
- Architecture design and implementation of Datalake which involves designing Data Ingestion, DAAS, Datamart, Reporting and Visualization layers (D3JS)
- Developed MIFID-2 regulatory reports on trades life cycle data using spark RDD and dataframe APIs
- Missing or erroneous data imputation using machine learning techniques like MLP and KNN classifications
- Risk analysis of market data by calculating profit & loss (PnL) and value at risk (VaR)
- Configuring data ingestion process using Flume and Kafka from multiple feeds and consolidate the data using rule-based validation and transformation process into single repository called datalake
- Developed housekeeping mechanism which clean-up unused data to maintain good health of cluster
- Working in agile software development model and drive the sprint cycle by breaking down the requirements, create estimate and distribute tasks among scrum team members
- JIRA creation, distribution and monitoring for each story and sub-tasks
- Regular code review of the completed JIRA tasks, creating GitHub branches, fixing code check-in conflicts and merging working codes to create final production ready branch for testing
- Creating test cases for all modules using unit testing methodologies
- Working knowledge on Cloudera CDH and Hortonworks HDP environment
- Experience in scheduling and managing the hadoop jobs using oozie
- Experience in staging log data from several application servers to HDFS using Flume, exporting structured data from RDBMS to HDFS using Sqoop and other data files used for common analysis
- Experience in taking training on statistical analysis, machine learning algorithms, Java, Scala, Python, Hadoop, Spark and other big data ecosystems

TECHNICAL PROFICIENCIES :

- **OS Platform:** CentOS, Ubuntu, RHEL and Windows Seven
- **IDE:** Eclipse, PyCharm and IntelliJ IDEA
- **Programming Language:** Java, Scala and Python
- **Web Server:** Apache Tomcat
- **Tracking tool:** JIRA
- **Code Repository tool:** GitHub and Bit Bucket
- **Build tool:** Jenkins and TeamCity
- **RDBMS Database:** IBM DB2**,** MySQL and Oracle
- **NoSql Database:** HBase
- **Big Data Ecosystem:** MapReduce, HDFS, Hive, Spark, Sqoop, Flume, Oozie and Kafka
- **Application Environment:** IBM Domino, IBM WebSphere Application Server and IBM Tivoli

PROFESSIONAL EXPERIENCE :

Société Generale Global Solution Centre -- 08/2017 - Present

Project: Datalake & Smart Automation
Technology Stack: Machine Learning, Deep Learning, Spark, Hadoop, Hive, Kafka, and Oracle
Languages: Python, Java and Scala
API & Libraries: D3JS, Scikit-Learn, TensorFlow, Keras, Pandas, Numpy, Seaborn and Bokeh
Role: Technical Lead (Data Science)
Responsibilities:
1. Statistical analysis, machine and deep learning algorithms to build smart automation systems along with building stock advisory application using Linear regression, SVR and LSTM prediction models with NLP to decipher social media and news feed sentiments
2. PDF, image and email document parsing and automating manual verification system using decision tree and other deep Learning algorithms. Bond and equity data analysis using visualization and time series models
3. Chabot implementation using NLP for department specific domain knowledge
4. Missing or erroneous data imputation using machine learning techniques like MLP and KNN classifications
5. Existing legacy architecture validation and proposal of new architecture with optimized data model along with big data stacks implementation
6. Architecture design and implementation of datalake which involves designing Data Ingestion, DAAS, Datamart, Reporting and Visualization layers (D3JS)
7. MIFID-2 regulatory reports on trades life cycle data using spark RDD and dataframe APIs

Key Achievements :

- Simultaneously worked on data science and datalake projects
- Designed datalake architecture and worked as a developer to implement the reporting modules
- Conducted internal trainings on statistical analysis, machine learning algorithms, python and scala

HCL Technologies Pvt. Ltd. -- 12/2015 - 08/2017

Project: Market Risk Analysis
Client: Deutsche Bank
Technology Stack: Spark, Hadoop, Hive, Scala, Java, Spring,Hibernate, Kafka, Sqoop and Flume
Languages: Python, Java and Scala
Role: Lead Data Engineer
Responsibilities:
1. Developing ETL process using spark applications to provide end to end data transformation and deep insight by calculating profit & loss (PnL) and value at risk (VaR) from raw market data
2. Integrating several data sources like Oracle database, HDFS raw files, Hive table data and data from NFS mount into Spark data abstraction called RDD and DataFrame where data are filtered and transformed and persisted back to hive tables to build meaningful insight using Historical Simulation and Monte-Carlo analysis in form of CSV reports and visual charts by business analysts
3. Using Spark streaming API for real time analysis of messaging data generated from vital business transactions across globe
4. Production bug fix and new release support along with admin team and client support for BAU
5. POCs on Spark MLlib , Apache Ignite, Akka, HBase, Oozie, Flume, Sqoop and Flink

Key Achievements :

- Driving the daily scrum stand-up and weekly training sessions
- Taking active participation in resource selection process and train them for required application stacks
- Conducting training on Big data technologies (Hadoop, Hive, Spark, Scala and Kafka)

IBM India Pvt. Ltd. -- 11/2010 - 12/2015

Project: IT Operations Analytics
Client: IGA (IBM Global Account)
Technology Stack: Hadoop, Hive, Oozie, Sqoop and Flume
Languages: Java and Scala
Cluster Infrastructure: 4 nodes IBM BigInsights cluster
Role: Data Engineer
Responsibilities:
1. Developing MapReduce application for transforming raw log files (semi-structured data) into structured format and persist it into hive partitions for further analysis using Hive (HQL and UDF)
2. Exporting semi-structured log files from different application servers to HDFS using Flume and exporting application related data from RDBMS to hive using Sqoop
3. Designing and maintaining hadoop job workflow using oozie
4. Complex data analysis from IBM Connections (Enterprise Social Networking) to identify top trending discussions and engagements on different technologies, ideas and improvement plans
5. Created solution documents and tools for the process using IBM Connections to make the work environment more social, business friendly, productive and innovative

Key Achievements :

- Among the few to get training and work for Big Data challenges for IBM operations analysis
- Provided training on hadoop fundamentals and developer modules for aspirants
- Created the first custom build hadoop cluster for testing purpose and hadoop blog for aspirants
- Developed IPT web tool for managing the IBM internal process which got recognized among the "Best Idea Implementation" category in 2013
- Won IBM's "Best Emerging Talent" award in 2013

EDUCATION AND TRAINING :

Academic Education:

- PGD (**Data Science**) from Manipal University, 2018
- PGD (**Computer Science**) from Annamalai University, 2010
- Graduation in Bio Science (Microbiology) from Calcutta University,2008
- 10+2 from Bholananda National Vidyalaya (C.B.S.E), 2005
- 10th from Satyanarayan Academy (C.B.S.E), 2003

Professional Training:

- **IBM ACSE** (Advanced Certificate in Software Engineering) certification from **IBM ACE**

Certifications:

- Statistics, Data Visualization, Machine learning, Deep Learning, Hadoop, Hive, Spark, Scala, Python and Kafka certifications from **Big Data University, Pluralsight, Datacamp and Coursera**

DECLARATION :

I hereby declare that information provided above is true to the best of my knowledge