https://www.github.com/ssomani7 https://www.linkedin.com/in/saurabhsomani7

Saurabh Somani

somani.saurabh7@gmail.com · +918856099399 · Latur, Maharashtra, India.

Summary

• Software Engineer with 2 years of industry experience with a knack for creative problem solving and collaboration.
• Solid foundation in computer science with strong competencies in data structures, algorithms and object-oriented programming. Strong fundamentals of *Data Engineering, Machine Learning, Distributed Computing and Software Development.*
• Experienced to work independently and in a globally diverse team with a fast-paced cross-functional development environment.
• As an individual you'll find me self-confident, result-oriented, creative, highly motivated and naturally passionate.

Availability

• Available to join on an ***IMMEDIATE basis.***

Education

Santa Clara University **Santa Clara, CA, USA**
M.S. Computer Science and Engineering | GPA: 3.50 / 4.00 *September 2017 – June 2019*
Related Coursework*: Big Data, Data Mining & Pattern Recognition, Advanced Database Systems, Machine Learning, Software Architecture, Design & Analysis of Algorithms, Object-Oriented Analysis Design & Programming, Operating Systems, Computer Networks.*

Vishwakarma Institute of Technology **Pune, India**
B. Tech. Computer Science and Engineering | GPA: 3.59 / 4.00 *August 2011 – May 2015*

Recent Project

Twitter streaming using Kafka, ElasticSearch & Avro Schema Registry *February 2020*
• Implemented ***end to end data pipeline*** for collecting live stream of tweets from Twitter. Created a Twitter developer account and an application in it. Using the Twitter API & Access tokens and keys from this app, user can collect live stream of tweets for the interested *topic of tweets* in real time.

• 4 different *Maven modules* for Java in backend illustrating 1) ***Idempotent*** Kafka Producer to get data from Twitter API into Kafka Topic. 2) ***Idempotent*** Kafka Consumer to get data from *Kafka* & storing it in *ElasticSearch* hosted in *bonsai.io cloud*. 3) A custom Java class for ***filtering Twitter tweets*** based on followers count & other features. 4) Performance improvement using ***Batching with Bulk Request Handling, Exception Handling for bad data and Logging.***

• Stored data locally in PostgreSQL with ***schema enforcement using Avro***. Tested REST proxy using Insomnia client.
• *Learning: Kafka Connect, Streams, Gson, REST Proxy, Twitter API | Language: Java8 | Frameworks: Apache Kafka, Avro, ElasticSearch, Docker, Insomnia REST Client, PostgreSQL, Postico.*

Professional Experience

Santa Clara University. **Santa Clara, CA, USA**
Media Systems Assistant III & Graduate Assistant *August 2018 – June 2019*
• Assisted in providing instructional technology and event support.
• Worked on "Video Duplication & Streaming" project using Apache Kafka and Apache Spark.
• Managed & Represented Department of Engineering at key university events.
• Used SAP-Concur to analyze & keep track of financial data for campus ministry.
• Official student member of Association of Computing Machinery (ACM).

Somani Distributors. **Latur, India**
Data Analyst *October 2016 – August 2017*
• Developed requirement specifications for pharmaceutical software.
• Data Analysis using SQL for generating purchase orders on repository stock.
• Successfully integrated the system with new patch updates, thus reducing the cost of new system by 50%.
• Developed new modules using Java at backend, JUnit for test cases & automated data to spreadsheet using Python3.

Cognizant. **Pune, India**

Programmer Analyst Trainee. *September 2015 – September 2016*

• Developed backend modules in Java for healthcare client using Spring MVC framework, Servlets and ESQL.
• Used Jenkins for continuous integration of code in the project.
• Performed ETL service for JSON, XML, CSV file formats and pushed to MongoDB using Node.js for server-side scripting.
• Lead major operations of the module for 2 weeks during the installation of new offshore development center. This helped in ensuring continuation of services without any extra cost, delay & reduction in the workload of new team.
• Used Scikit-Learn libraries for clustering modules of the project & improved performance using A/B testing.

Projects

Proof of Concepts with Apache Spark & Kafka *September 2019*

• Implemented Spark machine learning pipeline on AWS EMR for collaborative filtering to recommend users which online educational course they should take based on their viewing history. Target audience found using KMeans clustering over 2 billion data rows.
• Using Kafka & Spark Structured Streaming simulated the above models as real time events with a window size of 2 minutes.
• *Learning: RDD, MapReduce, Kafka, SparkML, Server Setup, Maven* | *Language: Java8* | *Frameworks: Spark, Kafka.*

Movie Lens Database – Hive Vs SQL *March 2019*

• Query implementation in Hive and SQL demonstrating the concepts of Schema-On-Read, Push-Predicate-Down, Hash-Joins, Indexing, Hadoop and MapReduce Working.
• *Learning:* Hive useful for analytical processing with huge data | *Language:* SQL | *Frameworks:* Apache Hive, Oracle 10g SQL.

Deep Learning Dashboard *September 2018*

• A dashboard web-app for interactive DNN model training, data management, performance analysis and visualization.
• *Learning:* DNN, Docker, Web-App Development, AWS | *Language:* Python3 | *Libraries:* Keras, Flask.

Yelp Data Application *March 2018*

• Java Swing based implementation of "Yelp Search Engine" with pagination, SQL database & MVC design pattern.
• *Learning:* B+ tree indexing, Push-Predicate-Down for narrow & fast searching| *Language:* Java8 | *Libraries:* Java Swing, JDBC.

Sentiment Analysis: Amazon Fine Food Reviews *November 2017*

• Sentiment Analysis on dataset with 560K+ reviews by 256K users with ratings and comments; used Word2Vec embedding to represent words; used linear, logistic regression, SVM to train the model with probabilistic outcomes for positive and negative sentiments.
• *Learning:* L1, L2 regularization, Word2Vec, Data Mining | *Language:* Python3 | *Libraries:* Scikit-learn, NLTK3.0, NumPy, Pandas.

Skills

Languages: Java (Proficient), Python3 (familiar).
Distributed Computing: Apache Spark, Kafka, Hive, SparkMLib, MapReduce, SparkSQL, PySpark, Hadoop.
Database Technologies: ElasticSearch, Apache Avro, MySQL, MongoDB, PostgreSQL, HDFS.
Operating Systems & Tools: Linux, macOS, Postman, Insomnia REST Client, Git, Eclipse, IntelliJ.
Cloud & Container Technologies: AWS, Docker.
Libraries: Scikit-Learn, NLTK3.0, Pandas, Matplotlib, Keras.
Web Technologies: Node.js, Django.

Hobbies/Personal Interests

• Watching European Soccer. Official membership holder of Liverpool F.C. for 2019/20 season.
• Cooking food, Travelling & Hiking.