Chandan Uppuluri

Email: chandan.uppuluri@gmail.com Phone: 682-712-3465
Web: http://chandan-u.github.io/ |GitHub: github.com/chandan-u |LinkedIn: linkedin.com/in/chandanu

EDUCATION

Masters in Data Science, INDIANA UNIVERSITY**,** Bloomington **Dec 2017**
sBachelor of Technology in Computer Science, GITAM UNIVERSITY, INDIA **Apr 2012**

SKILLS SUMMARY

Programming Languages: **Python**, **SQL**, Bash, **Scala, R**

Big Data Frameworks: **Apache Spark, Apache Airflow**, **postgres**, MySQL, **AWS Redshift, Delta.io**, ElasticSearch, AWS SageMaker, Kafka, Hadoop

Web/Visualization: **Django**, **FLASK**, HTML5, CSS3, RESTful API, SOAP, Swagger, D3.js, ggplot2, plotly, matplotlib, Tableau, **RShiny**

Platforms: Linux (Debian/rpm based), **DataBricks**, Paperspace, **AWS**

ML Tools: Tensorflow, Pandas, Keras, numpy, genism, nltk, scikit-learn

DevOps: **Docker**, **Ansible**

WORK EXPERIENCE

BitMex, San Francisco CA **Aug 2020 – Present**

Data Engineer (Fintech, AWS Redshift, Apache Airflow, AWS S3,, SQL, Python, ETL, Docker, Tableau, Salesforce, Sengrid)

• **Data Pipelines**: Developed ETL pipelines using python to populate Finance and Crypto Exchange data to:

A. Tableau Server : supports tableau dashboard
B. SalesCloud and Einstein Analytics (Salesforce): Supports Sales team.
C. SendGrid Email services: Using redshift data to send automated emails using Sendgrid API.

Zypmedia, San Francisco CA **Dec 2018 – Jul 2020**

Data Science Engineer (Adtech, Apache Spark, AWS Redshift, Apache Airflow, AWS S3, MySQL, SQL, Scala, Python, Linux, ETL, Databricks Delta, AWS SageMaker)

• **Data Pipelines**: Developed ETL pipelines using Apache Spark to populate:

A. Data Warehouse (AWS Redshift, Delta lake): used by analytics, BI teams, Marketing teams, Finance teams, for tracking ad performance, VCR, impression tracking, pacing, ad-hoc reports, custom client reports etc
B. Reporting DB (MySQL database): Which powers Dashboards for various clients, Advertisers, Media Companies.

• **Built End to End Streaming Solution:** Used Databricks Delta.io, Spark to empower advertisers with metrics in near real time **A.** In stream Joins
B. Deduplication
C. Single source of Truth and Unified Platform

• **Migration from AWS Redshift to Delta lakes (This made our solutions/pipelines move from ETL to ELT)**
• Use Statistics, Information Visualization and Data Mining, to look for anomalies and insights in data.
• Design, Schedule, monitor and orchestrate jobs using Apache Airflow.
• AWS SageMaker: Built a concept pipeline to train models to predict CTR.

Signet Accel Inc, Columbus OH **Jan 2018 – Oct 2018**

ETL Engineer (Healthcare, python, SQL, Apache Airflow, ElasticSearch, NLP, Postgres, Swagger, Django, Flask) • **ETL tool to de-identify of patient data (HIPAA)**
• **I**ntegrate data from various Hospitals into standard data models.
• Used NLP concepts (tf-idf etc) and ElasticSearch for text search and fuzzy matching medical data.
• Python and Apache Airflow for Orchestration, Scheduling and pipelines (Customized BaseOperators)
• **Continuous Integration and Continuous Delivery practices (**Ansible**)**
• Serving layer using Flask to support frontend services.
• **Lean ETL tool**, that could run on an instance as small as 2 cores and 10gb RAM.

World Well Being Project, UNIVERSITY OF PENNSYLVANIA **May 2016 - Aug 2016**

Data Science Research Assistant (Python, Data Engineering, nlp)

- Goal of the project is to predict emotions/empathy in social media data.
- Role: Implementing the data science concepts to reality.
- **Gathered required data** from various static (HDFS) and streaming data sources (twitter API) **to create data lakes**.
- **Cleaned unstructured data** (English & Arabic) for labelling on Amazon Mechanical Turk and model building.
- Visualized them using **word-clouds, LIWC (look for noise, insights)**
- Extracted features using n-grams, word-2vec, tf-idf
- **Used pandas, tensforlow, keras, scikit-learn API's for building machine learning models.**

School of Informatics and Computing, INDIANA UNIVERSITY **Jun 2016 - Sep 2016**

Instructor (Text Mining in Python Course, Capstone projects Teaching Assistant)

- **Developed online course for Indiana University**: NLP in python (INFO-I590) as part of Data Science online courses.
- **Capstone projects: Teaching Assistant –** Guiding undergrad CS students to build Apps, tools etc for their capstone projects.

HCL Technologies, India **Jul 2012 – Sep 2015**

Software Engineer, DATA (Telecom, middleware, Data Engineering)

- **Developed dashboards** with **Sqoop, Hadoop, Hive, Django, HTML, CSS, JQuery** to provide insights to client (T-Mobile) (POC)
- **Enabled backend integration** of T-Mobile MetroPCS merger through web services and message buses. (Enterprise Application Integration)
- **Automated work flows** which saved more than a hundred thousand dollars in a single quarter.

PUBLICATIONS

Building Customized Text Mining Tools via Shiny Framework: The Future of Data Visualization. (28th Modern Artificial Intelligence Cognitive Science Conference MAICS)

DATA ENGINEERING PROJECTS (Hobby/Academic)

Hadoop cluster from scratch (Spark, Linux, Hadoop, Hive, YARN, HDFS)

- Built a three-node cluster with one master and two slaves. (Ubuntu, CPU: 6, Mem: 16 GB, Disk: 60 GB).
- Configured Hadoop-2.7.4, Spark-2.2, SPARK on YARN with FAIR Scheduling mode. Didn't use any third-party distributions.

Client project: Analysis and Visualization of Trends in Translation Studies (live project)

- Created Data Lake from 1600+ publications and built pipelines to feed the Geo Spatial, Time Series, word clouds and topic analysis data visualizations and provide network graphs of co authorship networks.

Data Pipelining and Orchestration for Movie Recommendation System (Airflow, Spark, python3.5) (GitHub)

- Orchestrated End to End ETL pipeline using Apache Airflow. Data Transformations are done using Spark Dataframes.

Using Spark to parallelize Hyperparameter Tuning/Model Building (spark, Tensorflow, genism, scikit-learn) (GitHub)

- Used spark in a intuitive way to parallelize model learning using the concept of Grid Search.
- Improved accuracy by 9 percent in Predicting the growth of DJIA Stock index from top 25 news headlines
- Derived features such as n-grams, tf-idf, word2vec, doc2vec (word-embedding) for Naive Bayes, SGD, SVM, CNN and LSTM.

Document Similarity Algorithms: Analysis and Comparison (nltk, genism, deepdist, sparkML) **(**GitHub**)**

- Explored the disadvantages of Word2Vec in terms of scalability, concepts and disambiguation.
- Addressed the issues with alternative solutions such as LDA (Topic Modeling), Distributed Word2vec, Jaccard similarity etc.

Multi-Document Text Summarization using PageRank (nltk, networkx, scikit) **(**GitHub**)**

- Implemented simple extractive text summarization using graph techniques such as LexRank, degree centrality etc.
- Evaluated the algorithms using BLEU metrics against human generated summaries.

Storytelling through data (R, tidyr, ggplot2, plotly**)**

- "Safety and Neighborhoods: SFO", Wrote a story using San Francisco crime data to explain which neighborhood is safe in San Francisco. Used various visualizations and data analysis techniques.

Forecasting Housing Rental Demand (GitHub**)**

- Predict rental demand: Feature Extraction and Clustering improved accuracy from 70% to 75%.
- Built multi class classification models KNN, Random forests, Decision trees and XGBoost to predict housing demands in NY.