Gautam Dudeja

Email : gautam.dudeja@gatech.edu
Mobile : +1-334-332-3597

Skills

- **Languages**: Scala, Python, Java, R, SQL, Go **Frameworks**: Apache Spark, Flink, Beam, Dask, Pandas, Scikit-learn, PyTorch, Airflow

  Tools: AWS clouds (EMR, EKS, Lambda, Fargate, S3, DynamoDb, ElasticCache, Lambda, KMS, AWS Cloud Development Kit (CDK), AWS CloudFormation

  Hadoop: Hadoop MapReduce, Hdfs, Hive, Iceberg, Hudi, scoop, Oozie

Education

- **Georgia Institute of Technology** Atlanta, GA
  Master of Science in Analytics *Jan. 2021 – May 2023*
- **Auburn University** Auburn, Alabama
  Masters Science Computer Science and Computer Engineering; GPA: 4.00 *Aug. 2012 – July. 2014*
- **Auburn University** Auburn, Alabama
  Bachelor Science Electrical and Computer Engineering; GPA: 3.7 *Aug. 2008 – May. 2012*

Experience

- **Apple** San Francisco, CA
  Senior Software Engineer *Aug 2021 - Present*
  - Built a self serving Streaming and Batch framework to ingest data for different sources into data lake using Spark, Flink and Apache Beam and Iceberg Tables.
  - Build Monitoring and Observably library and defined KPI's for real time pipelines to monitor data quality using Grafana and Prometheus
  - Worked on a Unified data catalog which catalogs datasets from various stores like Kafka, Hive, S3, Elastic Search, Druid, Solr, Cassendra across Apple AML and also implemented discovery services which crawl these stores for updating the datasets
  - Built unified metadata catalog for managing data lineage, enable data discovery and governance using Datahub.
  - Built data maintenance and quality check tools for statistically profiling ingested data to track data quality thresholds.
  - Optimized models like XGBoost and SMOTE built using Spark MLib and SageMaker.
  - Work closely with stakeholders across the org from product engineers, data scientists, customer support, finance, and more, to build data pipelines that solve business needs
  - Tools: Apache Spark, Apache Beam, Dask, Kafka, Airflow, S3, EKS, Datahub
- **Paypal** San Francisco, CA
  Senior Software Engineer *Sep 2019 - Aug 2021*
  - Designed data transfer pipeline to get data from paypal events streams into PayPal central data lake using dataflow jobs in GCP.
  - Built platform on schema discovery, schema design and management service for data sources pulled across PayPal business units to build honey specific data lake.
  - Analytics environment based on docker and AWS, standardized the python and R dependencies. Wrote the core libraries that are shared by all data scientists.
  - Core service for all recommendation systems at Paypal, currently used on the paypal apps and throughout the content discovery process. Worked on both offline training and online serving.

- **Nielsen** Chicago, IL
  Data Engineer *Jan 201 and Sep 2019*
  - Designed and Developed distributed, fault-tolerant, scalable and reliable data ingestion platform for batch and stream processing in lines with lambda architecture for Nielsen customer data platform
  - Data ingestion process involved onboarding large customer base from multiple sources and billions of historical events which goes through hygiene, standardization and validation process.
  - Designed deduplication and Identity resolution algorithm for customer data from multiple sources to create one single 360 customer view using spark graph frame and calculate TF-IDF for text similarity.
  - Developed event stream and audience stream pipelines for clients to bring real time data into the platform using Nifi, Kafka and spark-streaming.
  - Designed elasticsearch indexes for clients to create search queries to target an audience for campaign execution via multiple channels like SMS, email or direct mail. Generate identity graphs from standardized customer data for identity resolution across sources to generate 360 view of customer.
  - Use statistical profiling for outlier detections and imputations to fill in missing attributes
  - Build ML-Ops infrastructure to support ML pipeline to build and deploy model artifacts in production

Publications

- Towards Thermal-Efficient Hadoop Clusters through Scheduling 2017 IEEE International Conference on Internet of Things (iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing (CPSCom) and IEEE Smart Data (SmartData) Year: 2017 — Conference Paper — Publisher: IEEE