Gautam Nayak

Email : gautam_snayak@outlook.com
Mobile : +91-636-107-1986

Summary

Results driven engineering professional comfortable switching between design and implementation of Big Data Systems, primarily in Healthcare and Insurance domains.Experience developing diverse solutions, methods for data analysis,data governance, collection and storage.Adept at picking up and standardizing modern and trending technologies into enterprise production scale applications.

Skills

Languages: Scala, Java
Technologies: Kafka, Spark, Flume, Hadoop (Cloudera), Solr, Web Services
Serialization: Avro, Parquet, JSON
DataStores: Cassandra, HBASE, Presto, Hive, Impala
CI/CD: Git, Jenkins, JIRA
Data Governance: Sentry, Navigator, Ranger
Message Oriented Middleware: AMQP Rabbit
Data Quality: Apache Griffin

Experience

- **Prudential - Cognizant** Newark, NJ/Bangalore
  Senior Data Engineer *Sept 2016 - Present*
  - **PruFast Track**:
    - Prudentials in-house-built Risk Assessment Mortality Model to fast track life insurance cases without examinations and invasive tests.
    - Developed features in the spark streaming application for fault tolerance such as Checkpoint Recovery, reduced the processing time for every application from few minutes to seconds.
    - Implemented the data-lake for orchestrating historical analysis of insurance applications which improved the efficiency of model by 30%.
    - Designed and built self-evolving flexible schema component in Avro to manage revisions over input data elements.
    - Utilized Lambda design to eliminate duplication of data over ingestion and transformation components.
  - **Disability Claims Prediction**:
    - Engine to predict the transition appropriateness of a short term claim to a Long term disability claim.
    - Designed the ingestion stack utilizing Sqoop and built the current data snapshot using Spark Batch.
    - Built a custom Data quality engine in Scala to translate custom DSL to Spark SQL.
    - Developed the business rules/filters processing and ML model implementation in Spark Scala from pilot to Production.
  - **Personalizations**:
    - Bulk data processing and injection service from Hadoop to Cassandra and provides a thin REST layer on top for serving offline computed data online.
    - Developed applications to serve personalized content to marketing campaigns and promotions over Eloqua.
  - **DataLake**: Built the datalake for PruFast Track, Disability Claims Prediction for offline analysis/ re-training models.
  - **Data Collection**: Internal platform repository for data from all Business Units.Built generic tools to transform data and provide it to Data scientists for analysis/research
  - **Recommendations**: Core service for all recommendation systems at Prudential, currently used on the homepage and throughout the content discovery process. Worked on both offline training and online serving.
  - **Content Discovery**: Improved content discovery by building new feature in Indexing service for Solr known as 'Infinite Depth Crawl' used extensively in various Business sites across Prudential.
  - **NoteBook Analytics**: Introduced Zeppelin which is extensively used across projects for Data discovery, analytics and Visualization on YARN through Spark using Scala, Python & R.

- **MD Anderson Cancer Center (Moon Shots) - Pwc** Houston, TX
  Senior Data Engineer *May 2014 - July 2016*
  - **Clinical Parsers** :
    - Designed & Built Parsers in Spark for processing large scale data related to clinical trials & Patient centric EHR systems such as ClinicStation and EPIC.
    - Redesigned and refactored existing Clinical parsersby migrating from Spark RDD's to Datasets to improve parsing performance by 40%
  - **DataLake**: Designed the complete datalake/Datastore platform in Hadoop & the ingestion pipeline which hosts the Clinical & Non Clinical data used for mining and building predictive models.
  - **Data Governance**: Architected the data governance platform in Hadoop using Sentry to mask/desensitize PHI data and its usage across the all the systems
  - **Rule Engine**: Built a horizontally scalable rule engine in Scala used in Dashboards for Researchers to mask/filter the data.Same rules were also used to publish workbooks in Tableau for visualizations
  - **MessageQueues**: Pub-sub application to consume high volume EPIC data from EHR systems for various patient centric domains
  - **Workflows**: Oozie an open source workflow framework to create and manage data pipelines leveraging reusables patterns to expedite developer productivity.
  - **cTAKES**: cTAKES an NLP framework for extraction of information from Electronic medical records. Modified a stage pipeline to annotate specific set of terms and improve the overall scoring confidence

- **Huawei - Mindtree** Bangalore, IN
  Senior Software Engineer *July 2011 - April 2014*
  - **Content Delivery Platform**: Created pipelines used to ingest data related to mobile content using Flume.Developed log analysis application in Map Reduce which analyzed the portal traffic later used to understand user activity.
  - **USSD Content delivery**: Service over USSD which suggests content to the end user based on the initial input, developed using fuzzy match
  - **Portal Framework**: Enterprise Java Web application involving content delivery to various service operators of Huawei designed for 23 countries in Middle East and Europe.Built features into the platform related to transactions, content enhancements.

Education

- **SDM College of Engineering & Technology** Dharwad, India
  Bachelor of Engineering in Electronics and Communication (8.38/10.00) *Aug. 2007 – July. 2011*