GITTU GEORGE

276 Bowdoin St | Dorchester, MA 02122 | 2812269194 | george.gittu@gmail.com

Education

University of Houston
M.S. in Computer Science (GPA 3.8)

Mahatma Gandhi University
B.S. in Computer Science

Skills

- **Programming:** Proficient in Java, C++, Linux shell scripts, SQL; Experience in R, Scala
- **Hadoop /Big Data:** MapReduce, Spark, Cascading, Hive, Pig, Sqoop, Flume, Oozie, NIFI, Falcon, MongoDB
- **Tools:** Zeppelin Notebook, Jupyter, Hue, Ambari Views, Tableau, Rstudio, Eclipse, MicroStrategy

Projects

World/Europe Racial Bias:
This project is inspired from the data visualization featured in Washigton Post on how impicit racial bias vary across United States. The data from IAT is collected for years from 2003 - 2015 with a total of 430k data points. After the data munging, map is drawn for both Europe and World to visualize these datapoints to see how implicit racial bias vary across Europe and when considering the entire world.

Spinraza Analytics:
As part of preparing for the launch of Spinraza in US to enter new therapy area for children suffering from Spinal Muscular Atrophy (SMA). We build & deploy a Spinraza Launch and Patient Services Data Repository & Analytics solution intended to support data ingestion, data processing & data extraction for the Launch Team, Patient Services team, Marketing team. Spinraza the first treatment for SMA got FDA approval in Dec 2016.

23andMe GWAS Study:
23andMe app gives ability to search for SNPs within the 23andMe data-set. The user can search for SNPs within Gene Regions, by specific SNP identifier and by phenotype all within a specific significance threshold. The study, which involved more than 9,000 individuals genotyped through 23andMe's direct-to-consumer testing service, looked for genetic associations related to nearly two-dozen common traits. The 23andMe project Wrangled 109 files together into 1.6 billion records. All the queries on SNP, Gene and Phenotype returned with Sub Second query times.

Analysis for Security Threat using Big Data Technologies:
Created application for Tietronix Software Inc. using Big Data Technologies to monitor and report security threats. The application enabled the company to study network log files in parallel to quickly identify and report unauthorized access, or access patterns that do not match proper application usage. Additional benefits included add/edit/remove rules for monitoring, log file locations, live vs. stored data analysis etc

Pinterest Data Analytics:
Crawling the Pinterest website for data collection and designed a database to store the data in real time. Utilizing various data analytics and statistics tools/packages to explore the data to test certain hypothesis we have on Pinterest usage pattern and networking behavior to answer many interesting questions for both research and business purposes

FastCP:
FactCP is a custom build AWS boto to utilize new corporate 10gpbs internet2 science cloud data transfer node. it has become obvious that corporate need a mechanism to speed up transfers in and out of amazon cloud. Current boto does it but we need to push the transfer rates higher to utilize the new connection and filling the gaps which has to do with the combination of multi-part transmissions and client side encryption using KMS. With this boto we were able to push speed by 10x with client side and server side encrytion.

Working Experience

Data Engineer **Cambridge, MA**
Biogen *2/2015 – Present*

- Worked in Agile – Sprint methodologies to do requirements gathering, analysis and planning
- Prepared both high & low level design docs and functional specification document
- Created DB, collection, projection, aggregation, replication, sharding in MongoDB
- Developed Spring batches for running batch jobs and schedule them to handle both one-time load and incremental updates from MySQL to MongoDB using Quartz
- Played a key role in tuning the performance of R Shiny tables when rendering large datasets.
- Managed scalable Hadoop clusters including cluster designing, provisioning, custom configurations, monitoring and maintaining using Hortonworks Distribution Platform.
- Did POC that demonstrated Hadoop tool integration into company's existing enterprise architecture
- Wrangled large datasets with PII information to use in data lake

- Handled various file formats such as Parquet, ORC, Avro, RCFile and JSON
- Designed and scheduled Oozie workflow for handling import and cleaning of data from various data sources using Sqoop, Pig, Hive and Shell script
- Developed UDFs for Hive using Java
- Created Hive tables to store the processed results in a tabular format and wrote Hive scripts to transform and aggregate data for running Tableau to generate report
- Built Zeppelin notebook and jupyter for data scientists, actively involved in further analysis in spark to design and implement extensible ETL framework for scalable machine learning pipelines

Teaching Experience

Teaching Assistant **Houston, TX**
University of Houston *01/2014-12/2014*

- Computer Security and Integrity
- Business Data Communications

Professional Development:

- HDP Developer: Apache Spark Training I (Biogen onsite 4 day training session)
- Advanced R programming: Cousera MOOC - Completed
- HDP Analyst -Data Science: Self Paced Training - Completed
- HDP Essensials -Apache Hadoop: Self Paced Training - Completed
- HDP Administrator: Self Paced Training -Completed
- Statistics with R - Duke University : Cousera MOOC - In Progress