Amboori Akhil Venkatesh

Denton, TX. Ph: +1 (469) 420-0792

akhilvenkatesh70@gmail.com

www.linkedin.com/in/akhil-venkatesh-a-154b652b3

SUMMARY

Experienced Data Analyst with 7 years of experience extensive expertise in data visualization, programming, cloud computing, databases, ETL, A/B testing and machine learning. I am proficient in SQL where I use stored procedures, triggers and complex SQL queries for Data analysis and leveraging Tableau for insightful analysis facilitating dashboard-based decision-making for diverse stakeholders. My proficiency in Python programming spans the entire data science project lifecycle encompassing data acquiring, data cleaning, exploration, and modeling using Pandas. Additionally, I worked with NoSQL databases and big data tools such as MongoDB, Hadoop, Hive and Spark with hands-on experience in AWS, Azure, and Google Cloud technologies. With a proven record of accomplishment in delivering data-driven insights and solutions in retail industry and translating business requirements into technical solutions, designing KPI's and closely collaborating with cross functional teams and business stakeholders, I strive to give my best and leverage my skills for the growth of businesses.

SKILLS

Big Data Eco-system: HDFS, MapReduce, Spark, Yarn, Hive, Pig, HBase, Sqoop, Flume, Kafka, Oozie, Zookeeper, NiFi, Impala

Data Visualization: Tableau, PowerBI, Parabola, QlikView, Seaborn, ggplot2, plot Ly, matplotlib, pandas, NumPy

Programming: Python, SQL, Scala, R, Java, C, HTML, CSS

IDE: Jupyter Notebook, Google Collab, Visual Studio Code

Cloud: AWS (S3, RDS Redshift, EMR), GCP, Azure, Snowflake

Database: MySQL, PostgreSQL, MS SQL Server, Hive, AWS Redshift, MongoDB, Big Query

MS Office & ETL: Word, PowerPoint, Excel, Informatica MDM, PySpark, Apache Hadoop, Hive, Apache Spark, Apache Airflow, Kafka, Docker

Machine Learning: Classification, Clustering, Regression, Statistical Analysis, Dimensionality Reduction, Data Wrangling, Data Mining, Data Modeling, Interpretation

Agile: Jira, Confluence

EXPERIENCE

Consumers Energy, Michigan **Jan 2024 – Present**

Big Data Engineer

Project Description: Consumers Energy is embarking on a transformative data engineering project aimed at leveraging data-driven insights to enhance energy efficiency, optimize operations, and improve customer satisfaction. The objective of this project is to develop a comprehensive data infrastructure and analytics platform that enables real-time monitoring, analysis, and optimization of energy consumption across the company's customer base.

Responsibilities

- Installed/Configured/Maintained Apache Hadoop clusters for application development and Hadoop tools like Hive, Pig, Zookeeper and Sqoop.
- Implemented Partitioning, Dynamic Partitions, Buckets in HIVE.
- Installed and Configured Sqoop to import and export the data into Hive from Relational databases.
- Developed MapReduce (YARN) jobs for cleaning, accessing, and validating the data.

- Administering large Hadoop environments build and support cluster set up, performance tuning and monitoring in an enterprise environment.
- Wrote MapReduce jobs using Pig Latin, Optimized the existing Hive and Pig Scripts.
- Designed, developed, and maintained data pipelines on GCP using tools like Apache Beam/Dataflow, Spark on Dataproc, and Apache Airflow for workflow orchestration, ensuring efficient data processing at scale.
- Used Python SAS to extract, transform & load source data from transaction systems, generate reports, insights, and key conclusions.
- Developed storytelling dashboards in Tableau Desktop and published on to Tableau Server which allow end users to understand the data on the fly with usage of quick filters for on-demand needed information.
- Developed reusable framework to be leveraged for future migrations that automates ETL from RDBMS systems to the Data Lake utilizing Spark Data Sources and Hive data objects.
- Developed product profiles using Pig and commodity UDFs.
- Load the data into HDFS from different sources like Oracle, DB2 by Sqoop and loaded into Hive tables.
- Designed and developed Pig Latin scripts and Pig command line transformations for data joins and custom processing of MapReduce outputs.
- Worked on google cloud platform (GCP) services like computer engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring and cloud deployment manager.
- Setup Alerting and monitoring using Stack driver in GCP.
- Design and implement large scale distributed solutions in AWS and GCP clouds.
- Monitoring the Hadoop cluster through MCS and working on NoSQL databases including HBase.
- Used Hive to analyze data ingested into HBase by using Hive-HBase integration and compute various metrics for reporting on the dashboard.
- Proficient in crafting complex HiveQL queries to extract required data from Hive tables and developing custom Hive UDFs.
- Automated the workflows using shell scripts to export data from databases into Hadoop.
- Configured Spark Streaming to receive real time data from Kafka and store the stream data to HDFS.

Environment: Hadoop YARN, Zookeeper, Spark 1.6, Spark Streaming, Spark SQL, Scala, Pig, Python, Hive, Sqoop Reduce, No SQL, HBase, Tableau, Java, AWS, GCP, Oracle 12c, Linux

ICICI Bank, India **Mar 2019 – DEC 2022**
Data Engineer / Hadoop Developer

- Design and implement scalable, fault-tolerant big data solutions using Hadoop and related technologies such as HDFS, MapReduce, Yarn, Hive, Pig, and Spark.
- Configure and manage Hadoop clusters using tools such as Cloudera Manager, Ambari, or Hortonworks Data Platform
- Develop and maintain data pipelines using tools like Apache NiFi, Apache Kafka, and Apache Storm.
- Build and maintain data warehousing solutions using Hive and Impala
- Optimize and improve the performance of Hadoop clusters by tuning parameters and implementing best practices.
- Collaborate with data scientists, data analysts, and other team members to support data-driven decision-making.
- Experience with big data processing and analysis frameworks such as Apache Spark, Storm, and Flink
- Experience with data integration and migration tools such as Apache NiFi, Apache Kafka, and Sqoop.
- Experience with cluster management and orchestration tools such as Cloudera Manager, Ambari, and Hortonworks Data Platform.
- Work with different data sources like HDFS, Hive and Teradata for Spark to process the data.
- Use Kafka a publish-subscribe messaging system by creating topics using consumers and producers to ingest data into the application for Spark to process the data and Configure Zookeeper to coordinate and support the distributed applications as it offers high throughput and availability with low latency.
- Configure Nginx to serve the static content of the web pages reducing the load on the web server for the static content.

- Write SQL queries to perform CRUD operations on PostgreSQL to save, store, update, and delete rows in tables using Play Slick.
- Create and update Jenkins jobs to develop pipelines to deploy the application in different environments like develop, QA and Production.

Environment: Spark, Zookeeper, SQL, Scala, Jenkins, Kafka, HBase, HDFS, Hive, Teradata, NiFi, Storm, Flink, HDFS, MapReduce, Yarn, Hive, Pig.

Yantra Software Solution, Hyderabad, IN **June 2016 – Jan 2019**

Data Analyst

- Created and analyzed business requirements to compose functional and implementable technical data solutions.
- Utilized Sqoop to ingest real-time data. Used analytics libraries Sci-Kit Learn, MLLIB and MLX tend.
- Extensively used Python's multiple data science packages like Pandas, NumPy, matplotlib, Seaborn, SciPy, Scikit-learn and NLTK.
- Performed Exploratory Data Analysis, trying to find trends and clusters.
- Worked on data that was a combination of unstructured and structured data from multiple sources and automated the cleaning using Python scripts.
- Developed Python Scripts to automate data validation and data cleaning process such as deduplicating and checking data consistency using Pandas and Apache Airflow.
- Wrote SQL queries to perform CRUD operations on PostgreSQL in tables using Play Slick.
- Tackled highly imbalanced Fraud dataset using under sampling with ensemble methods, oversampling and cost sensitive algorithms.
- Improved fraud prediction performance by using random forest and gradient boosting for feature selection with Python Scikit-learn.
- Implemented machine learning model (logistic regression) with Python Scikit- learn.
- Optimized algorithm with stochastic gradient descent algorithm Fine-tuned the algorithm parameter with manual tuning and automated tuning such as Bayesian Optimization.
- Developed a technical brief based on the business brief. This contains detailed steps and stages of developing and delivering the project including timelines.
- Measured the ROI based on the difference's pre-promo-post KPIs.
- Developed Hive queries that compared new incoming data against historic data. Built tables in Hive to store large volumes of data.
- Wrote the data validation SAS codes with the help of Univariate, Frequency procedures.
- Summarizing the data at customer level by joining the datasets of customer transaction, dimension and from 3rd party sources.
- Used HBase as the database to store application data, as HBase offers features like high scalability, distributed NoSQL, column oriented and real-time data querying to name a few.
- Utilized Kafka, a publish-subscribe messaging system by creating topics using consumers and producers to ingest data into the application for Spark to process the data and create Kafka topics for application and system logs.
- Developed Hive queries that compared new incoming data against historic data. Built tables in Hive to store large volumes of data.

Environment: Spark, Hadoop, AWS, SAS Enterprise Guide, SAS/MACROS, SAS/ACCESS, SAS/STAT, SAS/SQL, ORACLE, MS-OFFICE, Python (scikit-learn, pandas, NumPy), Machine Learning (logistic regression), Gradient Descent algorithm, Bayesian optimization, Tableau.

Education:

Bachelor of Commerce, Emeralds Degree college, IN.

Master of Science in Advanced data Analytics, University of North Texas, Texas.