Hema Latha

Atlanta, Georgia | Hemalathaa205@gmail.com | Phone: +1 4705084989 | Linkedin

SUMMARY

- 8+ years of total IT experience across a range of sectors, this includes hands-on experience in Big Data Analytics and development.
- Proven ability to use **Kafka** and **Spark** Streaming to gather, process, and aggregate significant quantities of streaming data.
- Knowledge of the **Star Schema** and **Snowflake** Schema Methodologies for creating Data Marts.
- Extensive experience in business intelligence software, including **Tableau** and **PowerBI**.
- Knowledge in designing and building Spark apps using **Python** to evaluate Spark's performance against **Hive**.
- Practical experience with distributed application architecture, software as a service, event-driven architecture, and service-oriented architecture (SOA) (SAS).
- Practical experience with the AWS family of services, including Amazon EC2, Amazon S3, Amazon RDS, VPC, IAM, Amazon Elastic Load Balancing, Auto Scaling, Cloud Front, CloudWatch, SNS, SES, and SQS.
- Working knowledge of the Hadoop ecosystem linked with the AWS Cloud platform and its many services, including Amazon EC2 instances, S3 buckets, and RedShift.
- Solid working knowledge of Azure Cloud Platform services such as Azure Data Factory (ADF), Azure Data Lake, Azure Blob Storage, Azure SQL Analytics, and Azure Databricks.
- Solid work history with cutting-edge technologies like Spark, Kafka, and Spark streaming.
- Collaborated throughout the business with cross-functional teams to construct and create proof of concept for enterprise Data Lake environments including MAPR, CLOUDERA, HORTONWORKS, AWS, and AZURE.
- Significant expertise in data analysis utilizing Drill, Pig Latin, HIVE, and Impala.
- Knowledge of building MAPREDUCE Java programs for preprocessing and cleaning up data.
- Solid comprehension of the Hadoop architecture and its numerous elements, including HDFS, Job Tracker, Task Tracker, Name Node, Data Node, Resource Manager, and Node Manager.
- Solid working knowledge of the integration of HBase/MapRDB with Hive.
- A thorough grasp and familiarity with NOSQL databases like Cassandra and HBase.
- Skilled in using Spark Data Frames and Python to translate Hive/SQL queries into Spark transformations.

AREAS OF EXPERTISE

Hadoop/Big Data	HDFS, MapReduce, Spark, Yarn, Kafka, PIG, HIVE, Sqoop, Storm, Oozie, Impala, HBase, Hue, Zookeeper.
Programming Languages	Python, R, Java PL/SQL, HiveQL, Scala, SQL
Development Tools	Eclipse, SVN, Git, Ant, Maven, SOAP UI
Databases	Greenplum, Oracle 11g/10g/9i, Teradata, Snowflake, MS SQL
No SQL Databases	Apache HBase, DynamoDB, Mongo DB
Hadoop Distributed platforms	Hortonworks, Cloudera.

PROJECT DETAILS

Pilot - Atlanta, GA **April 2021 - Till date**
Senior Data Engineer
Responsibilities:

- Collaborated with Business Analysts, Engineers across departments to gather business requirements, and identify workable items for development.
- Selected and generated data into CSV files and stored them into AWS S3 by using AWS EC2 and then structured and stored in AWS Redshift.
- Hands on experience working with AWS EMR, EC2, S3, Redshift, DynamoDB, lambda, Athena and Glue.
- Using Spark context, Spark-SQL, Data Frames, RDDs, and Memory optimization to investigate ways to make the existing Hadoop algorithms run faster and more efficiently.
- Experienced in designing and deployment of Hadoop cluster and different Big Data analytic tools including Pig, Hive, HBase, Oozie, Sqoop, Kafka, Spark with Cloudera distribution. Worked on Cloudera distribution and deployed on AWS EC2 Instances.
- Worked on integrating Apache Kafka with Spark Streaming process to consume data from external REST APIs and run custom functions.
- Involved in performance tuning of Spark jobs using Cache and using complete advantage of cluster environment.
- Configured, supported, and maintained all networks, firewall, storage, load balancers, operating systems, and software in AWS EC2.
- Implemented the use of Amazon EMR for Big Data processing among a Hadoop Cluster of virtual servers on Amazon related EC2 and S3.
- Involved in designing and deploying multi-tier applications using all the AWS services like (EC2, Route53, S3, RDS, Dynamo DB, SNS, SQS, IAM, Cloud formation) focusing on high-availability, fault tolerance, and auto-scaling in AWS Cloud Formation.
- Supporting continuous storage in AWS using Elastic Block Storage, S3, Glacier. Created Volumes and configured Snapshots for EC2 instances.
- Worked on ETL Migration services by developing and deploying AWS Lambda functions for generating a serverless data pipeline which can be written to Glue Catalog and can be queried from Athena.
- Creating S3 buckets also managing policies for S3 buckets and Utilized S3 bucket and Glacier for storage and backup on AWS.
- Worked on AWS hosted Databricks environment and used spark structured streaming to consume the data from Kafka topics in real time and perform merge operations on delta lake tables.
- Used Airflow as scheduling and orchestration tool of our data pipelines.
- Design, Development, Implementation ETL process to support CDC- Change Data Capture on Databricks platform.
- Used Snowflake extensively to do the ETL operations and imported the data from Snowflake to S3 and S3 to Snowflake.

Environment: Hadoop, Hive, AWS, MapReduce, Sqoop, Kafka, Spark, Yarn, Pig, PySpark, Shell Scripting, HBase, Scala, Cloudera, JUnit, Soap, Python, Teradata, MySQL.

Charter Communications - Alpharetta, Georgia **Sep 2018 - Mar 2021**
Data Engineer
Responsibilities:

- Hands-on experience in **Azure** Cloud Services (PaaS & IaaS), Azure Synapse Analytics, SQL Azure, Data Factory, Azure Analysis services, Application Insights, Azure Monitoring, Key Vault, Azure Data Lake.
- Used Azure Databricks to run Spark applications and triggered them from ADF using link services.
- Worked on performance tuning of spark applications and different optimization techniques to run query faster.
- Extracted the data from Teradata into HDFS/Databases/Dashboards using SPARK STREAMING.
- Used Azure Devops Pipelines for deploying new versioned applications and new ADF pipelines to the Production Environment.
- Created Delta Live tables as a streaming sink and created periodic aggregated stats on top of Delta Live table.
- Used Terraform to deploy Azure logic apps infrastructure in support of Azure Data Factory.
- Worked on Snowflake Data Lake, creating tables, pipes and stages and applying transformations.
- Designed the ETL process and created the high-level design document including the logical data flows, source data extraction process, the database staging, and the extract creation.
- Used Spark MLIB to predict Customer Demand for certain products for long-weekend sales and created score for high demand products which helped to manage store inventory to accommodate the customer's demand.
- Used Pyspark for data ingestion and perform complex transformations.
- Worked on importing data from MYSQL database to HDFS and vice-versa using SQOOP.
- Responsible for developing Kafka Producers and Consumers from scratch as per the requirement specifications.
- Developed Spark code and Spark-SQL/Streaming for faster testing and processing of data.
- Highly skilled in integrating Kafka with Spark streaming for high-speed data processing.
- Used Spark Data frames, Spark-SQL extensively to build multiple ETL pipelines.
- Converted RDD's to data frames to improve the performance and optimization using in-memory procedures with Spark Context, Spark-SQL, Data Frame, and Pair RDD's.
- Performance tuning using Partitioning and Bucketing of Hive tables.
- Optimized Java micro services and integrated Data sources Hive and Hbase, reduced latency by 27%.
- Extract Transform and Load data from sources Systems to Azure Data Storage services using a combination of Azure Data factory, T-SQL, Spark SQL, and U-SQL Azure Data Lake Analytics. Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in Azure Databricks.
- Implemented and deployed MLIB model using spark to predict next location of move.

Environment: Hadoop, Azure, HDFS, MapReduce, Spark, Pig, Hive, Impala, Sqoop, Kafka, HBase, Airflow, Tableau, Python, PL/SQL, Snowflake, Teradata, Linux shell scripting, Pyspark, PyCharm, Soap UI, Eclipse, Jenkins, Jira

Carvana - Saint Louis, MO **Oct 2016 - Sep 2018**
Data Engineer
Responsibilities:

- Worked extensively on designing and building of scalable flexible data solutions around batch, low latency, search, and real time data processing requirements using Spark, Kafka, HBase, Elastic search and Hadoop Eco-systems.
- Worked on the design and implementation of real time streaming ingestion using Kafka, and Spark Streaming.
- Work related to downloading Big Query data into pandas or Spark data frames for advanced ETL capabilities.
- Integrated UI components with backend data systems, including Spark, Kafka, HBase, and Elasticsearch to enable seamless data interaction and real-time updates.
- Implemented data visualization libraries and frameworks, such as D3.js and High charts, to create dynamic and informative data visualizations for analysis and reporting.
- Designed and developed UI components for real-time streaming data ingestion and processing using Spark Streaming and Kafka, providing live updates and insights.
- Utilized UI frameworks to create dashboards and reports for monitoring and analyzing data pipelines, including status, performance and error handling.
- Implemented user authentication and authorization mechanisms in the UI using Spring Security or similar frameworks, ensuring secure access to data engineering applications.
- Worked extensively in writing Kafka Producers to ingest data into Kafka topics using Java 8.
- Utilized Apache Hadoop by Hortonworks to monitor and manage the Hadoop Cluster.
- Completed data extraction, aggregation and analysis in HDFS by using PySpark and store the data needed to Hive.
- Working with DBA to design reports for DB replica latency trends, analyzing the transaction logs to find the root cause of the issues.
- Worked on Data ingestion to Kafka and Processing and storing the data Using Spark Streaming.

Environment: Hadoop (HDFS/Horton Works), Spark, Spark-SQL, Spark-Streaming, Scala, Kafka, Pig, Hive, Oozie, GCP, Linux, Splunk, Elastic search

Amdocs - India **May 2013 - June 2016**
SQL Developer
Responsibilities:

- Gathered business requirements, definition and design of the data sourcing, worked with the data warehouse architect on the development of logical data models.
- Ability to query data in a data warehouse and prepare data for reporting and insights automation needs.
- Extract, cleanse, and combine data from multiple sources and systems using R and Python Programming.
- Perform exploratory and targeted analyses, with a wide variety of statistical methods including cluster, regression, decision tree/random forest, time series using Python Programming.
- Built reports and dashboards to monitor KPIs (Key Performance Indicators) to understand drivers of KPI changes.
- Performed Regression testing for Golden Test Cases from State (end to end test cases) and automated the process using python scripts.
- Designed and executed analytic projects, generated insights to support business decisions using advanced analytical and visualization techniques such as descriptive, predictive, and prescriptive analytics.

- Generated graphs and reports using ggplot package in RStudio for analytical models. Developed and implemented R and Shiny application which showcases machine learning for business forecasting.
- Performed K-means clustering, Regression and Decision Trees in R. Worked on data cleaning and reshaping, generated segmented subsets using NumPy and Pandas in Python.
- Used Python NumPy, Pandas to perform data cleaning and data transformation activities.
- Scheduled data refresh on Tableau Server for weekly and monthly increments based on business change to ensure that the views and dashboards were displaying the changed data accurately.