NAGA PRIYANKA MUNNANGI

DATA ENGINEER

MAIL: mnagapriyanka1@gmail.com

PHONE: +1 (269) 2204591

SUMMARY:

- **8+** years of diversified experience in Software Design & Development.
- Experienced as Data Engineer solving business use cases for several clients.
- Experienced in the field of software with expertise in backend applications.
- Strong understanding of Distributed systems design, HDFS architecture, internal working details of MapReduce and Spark processing frameworks.
- Solid experience developing Spark Applications for performing highly scalable data transformations using RDD, Data frame, Spark-SQL, and Spark Streaming.
- Experienced in MVC and Microservices Architecture with Spring Boot and Docker, Swamp.
- Expertise in using Docker and setting up ELK with Docker and Docker-Compose. Actively involved in deployments on Docker using Kubernetes.
- Experience of developing applications with Model View Architecture (MVC2) using Spring Framework and J2EE Design Patterns.
- Strong experience troubleshooting Spark failures and fine-tuning long running Spark applications.
- Strong experience working with various configurations of Spark like broadcast thresholds, increasing shuffle partitions, caching, repartitioning etc., to improve the performance of the jobs.
- Experience in using the cloud services like Amazon AWS EMR, S3, Lambda, Auto Scaling, Cloud Watch, EC2, Red shift and Athena.
- Configured Spark Streaming to receive real time data from Kafka and store the stream data to HDFS and process it using Spark and Scala.
- Experience working with Data Lake is a system or repository of data stored in its natural/raw format, usually object blobs or files.
- Worked on Spark Streaming and Structured Spark streaming including Kafka for real time data processing.
- Strong experience of operating with cloud environments such as EC2 and S3 of Amazon Web Services (AWS).
- Solid experience in using the various file formats like CSV, TSV, Parquet, ORC, JSON and AVRO.
- Experienced working with various Hadoop Distributions (Cloudera, Hortonworks, Amazon EMR) to fully implement and leverage various Hadoop services.
- In depth knowledge on import/export of data from Databases using Sqoop.
- Well versed in writing complex hive queries using analytical functions.
- Knowledge in writing custom UDF's in Hive to support custom business requirements.
- Experienced in working with structured data using HiveQL, join operations, writing custom UDFs and optimizing Hive queries.
- Worked with Data Lake is usually a single store of all enterprise data including raw copies of source system data and transformed data used for tasks such as reporting, visualization, advanced analytics, and machine learning.
- Experience on Migrating SQL database to Azure Data Lake, Azure data lake Analytics, Azure SQL Database, Data Bricks and Azure SQL Data warehouse and controlling and granting database access and Migrating On premise databases to Azure Data lake store using Azure Data factory.
- Strong expertise in building scalable applications using various programming languages (Java, Scala, and Python).
- Continuous Delivery pipeline deployment experience with Maven, Ant, Jenkins, and AWS.
- Proficient in Core Java concepts like Multi-threading, Collections and Exception Handling concepts.
- Migrated SQL database to Azure Data Lake, Azure data lake analytics, Azure SQL database, Data Bricks and Azure SQL Data warehouse and Managing and granting access to and migrating to Azure data lake on-site databases using Azure Data factory research to Azure Data lake store.

- Strong experience in working with Databases like Oracle, and MySQL, Teradata, Netezza and proficiency in writing complex SQL queries.
- Experienced in version control tools like SVN, GitHub and CVS.
- Experienced working with JIRA for project management, GIT for source code management, JENKINS for continuous integration and Crucible for code reviews.

TECHNICAL SKILLS:

Languages	Shell scripting, SQL, PL/SQL, Python, R, PySpark, Pig, Hive QL, Scala, Regular Expressions
Hadoop Distribution	Cloudera CDH, Horton Works HDP, Apache, AWS
Big Data Ecosystem	HDFS, MapReduce, Hive, Pig, Sqoop, Flume, Oozie, Zookeeper, Kafka, Cassandra, Apache Spark, Spark Streaming, HBase, Flume, Impala
Databases	Oracle 10g/11g/12c, SQL Server, MySQL, Cassandra, Teradata, PostgreSQL, MS Access, Snowflake, NoSQL Database (HBase, MongoDB).
Cloud Technologies	Amazon Web Services (AWS), Microsoft Azure
Version Control	GIT, GIT HUB
IDE & Tools, Design	Eclipse, Visual Studio, Net Beans, Junit, CI/CD, SQL Developer, MySQL, SQL Developer, Workbench, Tableau
Operating Systems	Windows 98, 2000, XP, Windows 7,10, Mac OS, Unix, Linux
Data Engineer/Big Data Tools/Cloud/ETL/Visualization/Other Tools	Databricks, Hadoop Distributed File System (HDFS), Hive, Pig, Sqoop, MapReduce, Spring Boot, Flume, YARN, Hortonworks, Cloudera, MLlib, Oozie, Zookeeper, etc. AWS, Azure Databricks, Azure Data Explorer, Azure HDInsight, Linux, Bash Shell, Unix, etc., Tableau, Power BI, SAS, Crystal Reports, Dashboard Design

EDUCATION:

Masters in Data Science
Western Michigan University **January 2017 – Apr 2018**

Bachelor in Computer Science & Engineering
Vignan's Nirula Institute of Technology & Science for Women **August 2009 – May 2013**

PROFESSIONAL EXPERIENCE:

FedEx Ground (Pittsburgh, PA) | Jan 2021 – till date
Role: Data Engineer

Responsibilities:

- Experience using Impala for data processing on top of HIVE for better utilization.
- Configured Spark Streaming to receive real time data from the Apache Kafka and store the stream data to DynamoDB using Scala.
- Developed Spark code using Scala and Spark-SQL for faster processing and testing.
- Worked on Spark SQL for joining multi hive tables and write them to a final hive table and stored them on S3.
- Created Spark jobs to do lighting speed analytics over the spark cluster.
- Evaluated Spark's performance vs Impala on transactional data. Used Spark transformations and aggregations to perform min, max and average on transactional data.
- Data sources are extracted, transformed and loaded to generate CSV data files with Python programming and SQL queries.
- Implemented Spark RDD transformations to Map business analysis and apply actions on top of transformations.
- Wrote various SQL, PLSQL queries and stored procedures for data retrieval.
- Designed ETL using Internal/External tables and store in parquet format for efficiency.

- Developed multiple Kafka Producers and Consumers from as per the software requirement specifications.
- Performed advanced procedures like text analytics and processing using the in-memory computing capabilities of Spark.
- Used Amazon Web Services (AWS) which include EC2, S3, Cloud Front, Elastic File System, RDS, VPC, Direct Connect, Route53, Cloud Watch, Cloud Trail, Cloud Formation, and IAM which allowed automated operations. Worked on Cloudera distribution and deployed on AWS EC2 Instances.
- Configured Spark streaming to get ongoing information from the Kafka and store the stream information to AWS.
- Developed, deployed and troubleshot the ETL Workflows using Hive, Pig and Sqoop.
- Optimized Hive QL/pig scripts by using execution engine like Tez, Spark.
- Developed end to end data processing pipelines that begin with receiving data using distributed messaging systems Kafka for persisting data into Cassandra.
- Experienced in migrating Hive QL into Impala to minimize query response time.
- Collected data using Spark Streaming from AWSS3 bucket in near-real- time and performs necessary Transformations and Aggregations to build the data model and persists the data in HDFS.
- Responsible in creating Hive tables, loading with data and writing Hive queries
- Worked on User Defined Functions in Hive to load the data from HDFS to run aggregation function on multiple rows.
- Executed Hadoop/Spark jobs on AWS EMR using programs and data is stored in S3 Buckets.
- Used Zookeeper to store offsets of messages consumed for a specific topic and partition by a specific Consumer Group in Kafka.
- Responsible in creating mappings and workflows to extract and load data from relational databases, flat file sources and legacy systems using Talend.
- Extracted files from MongoDB through Sqoop and placed in HDFS and processed.
- Setup data pipeline using in TDCH, Talend, Sqoop and PySpark on the basis on size of data loads
- Wrote Map Reduce jobs using Java API and Pig Latin.
- Performed querying of both managed and external tables created by Hive using Impala.
- Implemented Real time analytics on Cassandra data using thrift API.
- Designed columnar families in Cassandra and Ingested data from RDBMS, performed transformations and exported the data to Cassandra.
- Developed data ingestion pipeline into AWS S3 buckets using Nifi.
- Created external and permanent tables in Snowflake on the AWS data.
- Explored with Spark to improve the performance and optimization of the existing algorithms in Hadoop using Spark context, Spark-SQL, PostgreSQL, Scala, Data Frame, Impala, OpenShift, Talend, pair RDD's.
- Fetched and generated monthly reports, Visualization of those reports using Tableau.
- Used Oozie Workflow engine to run multiple Hive and Pig jobs.
- Developed Impala scripts for end user/analyst requirements for ad Hoc analysis.
- Developed Spark/Scala, Python for regular expression (regex) project in the Hadoop/Hive environment with Linux/Windows for big data resources.
- Continuous monitoring and managing the Hadoop cluster through Cloudera Manager.
- Used Spark API over Cloudera Hadoop YARN to perform analytics on data in Hive.
- Created numerous ODI interfaces and load into Snowflake DB. worked on Amazon Redshift for shifting all Data warehouses into one Data warehouse.
- Deploy and Troubleshoot ETL jobs that use SSIS packages.
- Responsible to store processed data into MongoDB.

Environment: Kafka, MapReduce, Sqoop, Oozie, Tableau, Spark, Impala, YARN, Hadoop, Cloudera, HDFS, Hive, Pig, Flume, HBase, AWS, Java, Python, Solr. JUnit, Scala, Talend, PL/SQL, Oracle 12c, Snowflake DB, MongoDB, Tez and agile methodologies

Northwell Health (Lake Success, NY) | Sep 2019 – Dec 2020
Role: Data Engineer

Responsibilities:

- Wrote a Data Bricks code and ADF pipeline with fully parameterized for efficient code management.
- Created and maintained SQL Server scheduled jobs, executing stored procedures for the purpose of extracting data from Oracle into SQL Server. Extensively used Tableau for customer marketing data visualization.
- Created Power BI reports and upgraded power pivot reports to Power BI.
- Developed a detailed project plan and helped manage the data conversion migration from the legacy system to the target snowflake database.
- Transformed business problems into Big Data solutions and define Big Data strategy and Roadmap. Installing, configuring, and maintaining Data Pipelines.
- Developed Data Bricks Python notebooks to Join, filter, pre-aggregate, and process the files stored in Azure data lake storage.
- Utilized Power Query in Power BI to Pivot and Un-pivot the data model for data cleansing and data massaging.
- Designed and maintained ADF pipelines with activities – Copy, Lookup, For Each, Get Metadata, Execute Pipeline, Stored Procedure, If condition, Web, Wait, Delete etc.
- Designed, developed, and tested dimensional data models using Star and Snowflake schema methodologies under the Kimball method.
- Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in in Azure Databricks.
- Implemented Copy activity, Custom Azure Data Factory Pipeline Activities.
- Designed and developed business intelligence dashboards, analytical reports and data visualizations using Power BI by creating multiple measures using DAX expressions for user groups.
- Developed Kafka producers and consumers efficient ingested data from various data sources.
- Responsible for wide-ranging data ingestion using Sqoop and HDFS commands.
- Accumulate ‘partitioned’ data in various storage formats like text, JSON, Parquet, etc. Involved in loading data from LINUX file system to HDFS.
- Primarily involved in Data Migration using SQL, SQL Azure, Azure Storage, and Azure Data Factory, SSIS, PowerShell.
- Designed the business requirement collection approach based on the project scope and SDLC methodology.
- Managed Azure Data Lakes (ADLS) and Data Lake Analytics and an understanding of how to integrate with other Azure Services. Knowledge of USQL.
- Created Pipelines in ADF using Linked Services/Datasets/Pipeline/to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
- Worked on tickets opened by users regarding various incidents, requests.
- Wrote production level Machine Learning classification models and ensemble classification models from scratch using Python and PySpark to predict binary values for certain attributes in certain time frame.
- 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.
- Used Apache Spark Data frames, Spark-SQL, Spark MLlib extensively and developing and designing POC's using Scala, Spark SQL and MLlib libraries.
- Designed both 3NF data models for OLTP systems and dimensional data models using star and snowflake Schemas.
- Developed various Mappings with the collection of all Sources, Targets, and Transformations using Informatica Designer.
- Creation of data aggregation and pipelining using Kafka and Storm.
- Used SQL Server Integrations Services (SSIS) for extraction, transformation, and loading data into target system from multiple sources.
- Involved in Unit Testing the code and provided the feedback to the developers. Performed Unit Testing of the application by using NUnit.
- Wrote research reports describing the experiment conducted, results, and findings and make strategic recommendations to technology, product, and senior management.
- Worked closely with regulatory delivery leads to ensure robustness in prop trading control frameworks using Hadoop, Python Jupyter Notebook, Hive and NoSQL.

- Wrote UNIX shell scripts to automate the jobs and scheduling Cron jobs for job automation using commands with Crontab.

Environment: Hadoop, Kafka, Spark, Sqoop, Snowflake, Spark SQL, Spark-Streaming, Hive, Scala, pig, NoSQL, Impala, Oozie, HBase, Zookeeper, Power BI, Azure, Data Bricks, Data Lake, Data Factory, Unix/Linux Shell Scripting, Python, PyCharm, Informatica, Linux, Shell Scripting, Informatica PowerCenter

Pentagon Federal Credit Union (Tysons Corner, VA) | Jun 2017 – Aug 2019
Role: Data Engineer

Responsibilities:

- Wrote various data normalization jobs for new data ingested into Redshift.
- Developed SSRS reports, SSIS packages to Extract, Transform and Load data from various source systems.
- Implemented and managed ETL solutions and automating operational processes.
- Optimized and tuned the Redshift environment, enabling queries to perform up to 100x faster for Tableau and SAS Visual Analytics.
- Advanced knowledge on Confidential Redshift and MPP database concepts.
- Migrated on premise database structure to Confidential Redshift data warehouse.
- Defined facts, dimensions and designed the data marts using the Ralph Kimball's Dimensional Data Mart modeling methodology using Erwin.
- Strong understanding of AWS components such as EC2 and S3.
- Implemented a Continuous Delivery pipeline with Docker, and Git Hub and AWS.
- Built performant, scalable ETL processes to load, cleanse and validate data.
- Participated in the full software development lifecycle with requirements, solution design, development, QA implementation, and product support using Scrum and other Agile methodologies.
- Compiled data from various sources to perform complex analysis for actionable results.
- Measured Efficiency of Hadoop/Hive environment ensuring SLA is met.
- Worked publishing interactive data visualizations dashboards, reports/workbooks on Tableau and SAS Visual Analytics.
- Worked on Big data on AWS cloud services i.e., EC2, S3, EMR and DynamoDB.
- Created Entity Relationship Diagrams (ERD), Functional diagrams, Data flow diagrams and enforced referential integrity constraints and created logical and physical models using Erwin.
- Created ad hoc queries and reports to support business decisions SQL Server Reporting Services (SSRS).
- Analyzed the system for new enhancements/functionalities and perform Impact analysis of the application for implementing ETL changes.
- Involved in the Forward Engineering of the logical models to generate the physical model using Erwin and generate Data Models using ERwin and subsequent deployment to Enterprise Data Warehouse.
- Collaborated with team members and stakeholders in design and development of data environment.
- Prepared associated documentation for specifications, requirements, and testing.
- Managed security groups on AWS, focusing on high-availability, fault-tolerance, and auto scaling using Terraform templates. Along with Continuous Integration and Continuous Deployment with AWS Lambda and AWS code pipeline.
- Created various complex SSIS/ETL packages to Extract, Transform and Load data.
- Was responsible for ETL and data validation using SQL Server Integration Services.
- Defined and deployed monitoring, metrics, and logging systems on AWS.
- Connected to Amazon Redshift through Tableau to extract live data for real time analysis.
- Used Hive SQL, Presto SQL and Spark SQL for ETL jobs and using the right technology for the job to get done.
- Analyzed the existing application programs and tune SQL queries using execution plan, query analyser, SQL Profiler and database engine tuning advisor to enhance performance.
- Optimized the TensorFlow Model for efficiency.

Environment: AWS, EC2, S3, SQL Server, Erwin, Oracle, Redshift, Informatica, RDS, NOSQL, MySQL, Dynamo DB, Docker, PostgreSQL, Tableau, Git Hub

NeoSOFT (Mumbai, India) | Aug 2013 – Nov 2016
Role: Data Engineer

Responsibilities:

- Participated in requirements sessions to gather requirements along with business analysts and product owners.
- Involved in Kafka and building use case relevant to our environment.
- Worked on implementation and maintenance of Cloudera Hadoop cluster.
- Pulled the data from data lake (HDFS) and massaging the data with various RDD transformations.
- Involved in building an information pipeline and performed analysis utilizing AWS stack (EMR, EC2, S3, RDS, Lambda, Glue, SQS, and Redshift).
- Responsible for developing data pipeline using flume, Sqoop and pig to extract the data from weblogs and store in HDFS.
- Developed Oozie workflow jobs to execute hive, Sqoop and MapReduce actions.
- Architected, Designed and Developed Business applications and Data marts for reporting.
- Imported the data from different sources like HDFS/HBase into Spark RDD and developed a data pipeline using Kafka and Storm to store data into HDFS.
- Collaborated with Business users for requirement gathering for building Tableau reports per business needs.
- Developed Pig Latin scripts for replacing the existing legacy process to the Hadoop and the data is fed to AWS S3.
- Developed continuous flow of data into HDFS from social feeds using Apache Storm Spouts and Bolts.
- Involved in loading data from Unix file system to HDFS.
- Implemented the Big Data solution using Hadoop, hive and Informatica to pull/load the data into the HDFS system.
- Objective of this project is to build a data lake as a cloud-based solution in AWS using Apache Spark.
- Installed and configured Hadoop Ecosystem components.
- Developed Spark code using Scala for faster testing and processing of data.
- Apache Hadoop installation & configuration of multiple nodes on AWS EC2 system.
- Created Hive External tables to stage data and then move the data from Staging to main tables.
- Documented the requirements including the available code which should be implemented using Spark, Hive, HDFS, HBase and Elastic Search.

Environment: Hadoop, YARN, HDFS, Spark, flume, AWS, Sqoop, pig, MapReduce, UNIX, Zookeeper HBase, Kafka, Scala, NoSQL, Cassandra, Elastic Search, Sqoop