**Nagendra Babu  
Big Data Engineer-ML**

Professional Summary:

* Having 10+ Years of IT Experience in Application development using various niche technologies like **Big Data, Machine Leaning and Apache SPARK (SQL API, Streaming API), Python, Scala, Java, Airflow, Autosys, Git, Hive, AWS, Azure Cloud Services**
* Excellent understanding of Hadoop, **Spark architecture and various ecosystem** components such as HDFS, Spark, Hive, HBase, Yarn, Spark SQL and Spark Streaming
* Experienced in working with **Datasets, Spark-SQL, Data Frames, RDD's**, handling large data frames, Spark in-Memory capabilities, transformations and other during ingestion process.
* Experience in various Big Data application phases like Data Ingestion, Data analytics and Data visualization.
* Experience in **data processing** like collecting, aggregating, moving from various sources.
* Hands on experience in developing **ETL data pipelines using pyspark** on AWS EMR
* Perform data cleaning, wrangling, mugging and merging to various different sources to make the data more meaningful
* Experience in converting SQL queries into **RDD/Data Frame transformations using Apache Spark using Python** and Java.
* Good working experience on different file formats (PARQUET, TEXTFILE, AVRO) and different compression codecs (GZIP, SNAPPY, LZO).
* Good working knowledge on **Spark Structured Streaming using Kafka and ETL** batch jobs on AWS infrastructure using EMR, S3, EC2 etc.
* Design, build and maintain **ML models** in production working closely with modeler’s and business stake holders
* Valuable experience on practical implementation of cloud-specific **AWS or Azure cloud** technologies including Amazon Cloud Services like Elastic Compute Cloud (EC2), Simple Storage Services (S3) and EMR
* Experienced in **trouble shooting spark jobs**.
* Knowledge of monitoring spark jobs using Spark UI
* Experienced in dealing with AWS S3 object storage from Spark
* Have hands on **NoSQL databases like HBase, DynamoDB, MongoDB** and Elastic Search.
* Understanding knowledge of Relational **Data Modelling and Data Warehousing** concepts
* Good knowledge on various scripting languages like **Linux/UNIX shell scripting** and **Python**.
* Understanding of software skills such as business analysis, development, maintenance, and software improvements (**SDLC**).
* Performing **Team Lead** Activities and Coordination with the team members and defining time estimations for deliverables of change requests, patches and upgrades to the application.
* Worked in **SCRUM** (Agile scrum model) team in delivering agreed user stories on time for every sprint.

Technical Skills:

| Languages | Python, Java, Scala, SQL |
| --- | --- |
| Big Data **Technologies** | Hadoop, AWS S3, Parquet, Hadoop Distribution: AWS EMR, SPARK SQL, Spark Streaming, Query Engine: Spark SQL, ElasticSearch, Hive, HBase, Yarn, Hue, Sqoop, Kafka |
| Cloud Platforms | Amazon Web Services (AWS Cloud IaaS), Azure. |
| ML Models | Data Preparation, Regression (Least Squares, Linear Regression, Logistic Regression), Time Series, Forecasting, Decision Support Algorithms (Decision Trees, Random Forest), Clustering, Optimization Techniques, XGBoost, SciPy, Numpy, Pandas, scikit-learn, scipy, native libraries, matplotlib |
| Data Sources | SQL, ORACLE, SYBASE, MySQL, DynamoDB, RDS, Postgres SQL, MongoDB, Teradata, AWS S3, File based |
| Version Control System | Git, SVN |
| Development / Build Tools | Eclipse, GIT, Maven, IntelliJ, PyCharm |
| **Operating systems** | UNIX, Mac and Windows Variants |
| **Methodologies** | Agile |

Education Summary

| **Degree** | **University** | **Specialization** |
| --- | --- | --- |
| Bachelor of Technology in Computer Science (B. Tech) - 2013 | JNTU Hyderabad, Telangana, India | Computer science (CSE) |

Certifications:

* AWS Certified Solutions Architect
* Google API Developer Certified

Professional Experience:

**Client - JP Morgan Chase & Co. Jan 2023 - Present**

**Project - Pricing Elasticity and Optimization (PEO)**

**Big Data Developer with ML (Software Engineer II)**

**Responsibilities:**

* Work closely with internal stakeholders such as business teams, product managers, engineering teams and partner data modelling, architecture teams.
* Implemented Spark using Python and utilizing Data frames and Spark SQL API for faster processing of data.
* Building pipelines in pyspark, tuning spark queries
* Work on Spark Context, Spark-SQL, Data Frame, Datasets, Spark YARN.
* Perform Data Cleaning, features scaling, features engineering using pandas and numpy packages in python.
* Perform operations on different data formats such as CSV, JSON, Parquet files from s3 and perform data analysis and validations
* Design, build and maintain scalable ML models in production environment and monitor them
* Build the python model optimization code and deploy to AWS Lambda as a API service to product users
* Build reusable components and deploy to all environments using Jenkins pipelines
* Performance improvements to existing implementation spark code and reduce run time.
* Implement new features to the current built in solution, this feature will allow the optimizer to take input subsidies to predict volumes and SVA and also by-pass optimization. This will allow business users to get pricing on products quickly
* Write unit test cases which ensures that all code meets quality standards
* Apply knowledge of machine learning model frameworks, algorithms, and tools for building ML solutions
* Contribute to the group’s knowledge base by finding new and valuable ways to approach problems and projects.
* Communicate the results with product owner’s team for taking best decisions on pricing.

**Environment:** Python, Spark, hive, HDFS, AWS, S3, EMR, Lambda, API Gateway, ML models - XGBoost, SciPy optimization (minimize), Terraform etc.

**JP Morgan Chase & Co. April 2020 - Dec 2022**

**Project - Consumer Analytics and Reporting Infrastructure (CARI)**

**Big Data Engineer (Software Engineer I)**

**Responsibilities:**

* Involved in the process of data acquisition, data pre-processing and exploring project in python/Scala/Java
* Coordinate with variety of teams to explore various data sources that best represent the project needs by having synchronous sprints in a distributed scrum environment which aids the flow of information and enhance communication across teams
* Involved in data migration work from onprem to aws cloud by building reusable data frameworks
* Involved in the loading of structured and unstructured data into HDFS
* Loaded data from Teradata, a relational database to HDFS on regular basis using Spark SQL API
* Build an enterprise-wide ETL Pipeline workflows with Batch Processing involving massive (~TB’s) amount of data Building data ingestion pipelines using spark on cloudera-CDH and AWS cloud
* Building history and incremental data ingestion pipelines through workflows
* Working on creating staging tables in Hive and load data to it
* Creating Autosys jobs to triggers data pipeline jobs
* Working on private and public cloud solutions such as AWS and JPMC internal cloud
* Spin up the AWS services such as EMR, S3, Glue tables by using the Terraform scripts
* Migrated data existing in Hadoop cluster into spark and used Spark SQL and Scala to perform actions on the data.
* Creating S3 buckets and managing policies for S3 buckets and utilized S3 bucket and Glacier for storage and backup AWS.
* Experienced with Spark Context, Spark-SQL, Data Frame, Datasets, Spark YARN.
* Worked with SCRUM team in delivering agreed user stories on time for every sprint.

**Environment:** Data Storage: CARI Hive Tables, Storage Format Parquet, CSV, tables, HDFS, HBASE, CDH/AWS, Alteryx Designer, Spark, AWS S3, EMR, Athena, Glue, Airflow, Autosys.

**S&P Global Feb 2019 - April 2020**

**Project - IDF REAL TIME REPLICATION ENGINE**

**Big Data Engineer**

**Responsibilities:**

* Build reusable frameworks which will execute the transform logic and load data into Data Lake or data warehouse by using Spark, AWS EMR, S3, HDFS, Hive, S3 etc
* Responsible for ingesting large volumes of data to Kafka.
* Wrote Kafka producers to stream the data from external rest APIs to Kafka topics.
* Explore discrete data sources in coordination with other teams to transform data at hand to create data lake Building streaming data pipelines using spark structured streaming and spark batch process on AWS EMR
* Good understanding of spark eco system and execution of jobs on yarn cluster
* Hands on experience of parse Kafka messages using spark RDD and Data Frames.
* Hands on experience on Spark's RDD, DataFrame and Dataset API’s
* Experience of dealing with AWS S3 object storage from Spark.
* Experience in trouble shooting spark jobs.
* Experience in performance tuning of Spark jobs.
* Experience of dealing with persistent Incoming Stream Data into s3 data lake
* Store Kafka offsets into External Store oracle
* Store Kafka offsets into spark inbuilt checkpoint Directory
* Collecting streaming metrics using spark streaming listener
* Alert Mechanism for Failures and Restart of streaming application

**Environment:** Data Storage: AWS S3, Storage Format: Parquet, Hadoop Distribution: AWS EMR, Change Data Capture: Oracle Golden gate with Big Data Connector, Streaming: Kafka, Spark SS (Scala), Meta store: Oracle, Query Engine: Spark SQL, Elastic Search.

**Tata Consultancy Services Feb 2016 – Jan 2019**

**Project - Cambia Health Care Member Communication Service**

**Big Data Engineer / Spark developer**

**Responsibilities:**

* Understand the business requirements from the stake holders and extract data from a range of sources, such as APIs, non/relational databases, XML, JSON, CSV files, HDFS and convert it into a single format for standardized processing
* Cleansing the data received from data source to remove duplicate and irrelevant observations, fix structural errors, filter unwanted outliers and handle missing data using algorithms
* Building ETL data pipelines using pyspark on AWS EMR by using Spark's RDD API and Spark's Data frame
* Hands on experience of XML processing using python spark
* Experience in configuring EMR clusters on AWS
* Experience and good understanding of Apache Spark Data sources API.
* Experience of dealing with AWS S3 object storage from Spark.
* Experience in trouble shooting spark jobs.
* Performance tuning of Spark jobs.
* Experienced in validating the data and finding the missing records by building the SQL queries.
* Enhancements to existing ETL batch processing frameworks as per business requirements and needs

**Environment:** Python, Spark2.0, Yarn, Hive, Parquet, ORC file formats, HDFS, AWS (EMR, RDS, S3), Jenkins, Code commit, Git, Eclipse IDE and Airflow

**Tata Consultancy Services Aug 2015 – Jan 2016**

**Project - Digital Enterprise Solutions and Services**

**Big data / Cloud Developer**

**Responsibilities:**

* Developed Java classes as Business Logic and Data Access Layer to communicate with the Rest API’s.
* Created and maintained the configuration of the Java MVC Framework module services in order to access the unified API's of these modules.
* Exposed few API's like SOAP, UDDI, WSDL through Web Services and wrote Web-services to expose the business methods to external services.
* Good Knowledge and experience with RESTful API’s.
* Performed a POC by making use of High Availability (HA) Multi AZ deployment supported by AWS Database (RDS) providing a secure access between EC2 Instances and RDS using VPCs.
* Worked with cloud services like Confidential Web Services(AWS) and involving in ETL, Data integration and Migration.
* Performed a POC to access the data (Video file) present in S3 directly and by using AWS Cloud Front distribution from different EDGE locations which improves accessibility.
* Used Confidential cloud-watch to monitor and track resources on AWS.
* Performed a POC by deploying a Stand Alone On-premises Web based application on EC2 servers.
* Used confidential cloud-watch to monitor and track resources on AWS.
* Gained hands-on knowledge on the following AWS services:
* Compute Services - Amazon EC2, EMR
* Storage Services - EBS Volume, S3
* Database - RDS, DynamoDB
* Security & Identity – IAM
* Application Services – SES, SQS, SWF and SNS
* Distribution Services - Cloud Front
* Networking - VPC, Route53
* Supporting Services – ELB, Auto scaling
* Management Tools – Cloud Watch, Cloud Formation

**Environment:** Java Servlet (JSP) web application, HTML, Linux, AWS (EC2, EBS, RDS, S3, IAM, SQS, SNS, Cloud Front, VPC, Route53, Cloud Watch), Jenkins, Code commit, Git and Airflow

**Rational Technologies May 2013 – July 2015**

**Project – Enterprise Data Lake (EDP)**

**Hadoop Developer**

**Responsibilities:**

* Responsible for loading customer's data and event logs into HBase using Java API.
* Created HBase tables to store variable data formats of input data coming from different portfolios
* Involved in adding huge volumes of data in rows and columns to store data in HBase
* Responsible for architecting Hadoop clusters with CDH4 on CentOS, managing with Cloudera Manager.
* Involved in initiating and successfully completing Proof of Concept on Flume for Pre-Processing,
* Used Flume to collect the log data from different resources and transfer the data type to Hive tables using different SerDes to store in JSON, XML and Sequence file formats.
* Used Hive to find correlations between customer's browser logs in different sites and analyzed them.
* End-to-end performance tuning of Hadoop clusters and Hadoop MapReduce routines against very large data sets.
* Created and maintained Technical documentation for launching Hadoop Clusters and for executing Hive queries and Pig Scripts.
* Created User accounts and given the users the access to the Hadoop Cluster.
* Developed Pig Latin scripts to extract the data from the web server output files to load into HDFS
* Developed the Pig UDF's to pre-process the data for analysis.
* Loaded files to Hive and HDFS from MongoDB Solr.
* Monitored Hadoop cluster job performance and performed capacity planning and managed nodes on Hadoop cluster.
* Responsible for using autosys to control workflow.

**Environment:** Hadoop 2.0, HDFS, Pig 0.11, Hive 0.12.0, MapReduce 2.5.2, Sqoop, LINUX, Kafka 0.8.1, HBase 0.94.6, CDH4, Autosys.