**Karthik Gorrepati**

**Sr. Data Engineer**

**Contact:** +**1 510-854-1314**

# EXPERIENCE SUMMARY

* **13+ Years** of IT experience with **5+** years of Experience in **Bigdata** Technologies like **Hadoop, Spark, Kafka, Spark Streaming, Hive, Scala, MapReduce, JAVA, ELK** and **strong** experience in C, C++, VC++ on Windows and Unix.
* Exposure in Hadoop technologies such as **HDFS, Hive, SPARK, SPARK SQL, Data Frames, Scala, Python, Sqoop**, **Kafka, Spark Streaming, Oozie, Flume, ELK.**
* Implemented performance tuning techniques for Hive queries. Expert in working with Hive tables, data distribution by implementing partitioning and bucketing, writing and optimizing the HiveQL queries.
* Hands on experience in handling different file formats like PARQUET, Avro, ORC Files, CSV, JSON, log.
* Good Experience with **NoSQL** Databases like **MongoDB** and **HBASE**.
* Has hands-on Experience in design & developing the application independently using MVC design pattern along with Singleton, Factory, Publish & Subscribe, façade, State and various design patterns.
* End to End Experience in handling the applications independently.
* Has Good Experience in working with Business Analysts to transform use cases into Hadoop centric implementation.
* Has extensive experience working with EMR, Cloudera and Hortonworks Distribution
* Always been successful in completing the projects on time while meeting strict timelines and requirements. This has enabled me in establishing strategic relationships with key clients and prospects.
* Excellent Ability to learn new technologies as project demands
* Very **good debugging, analytical and logical skills**.
* Excellent **presentation, documentation and communication skills**.
* Highly motivated and can work independently or part of a team
* Have excellent knowledge of **Software Development Life Cycle and Agile - SCRUM Methodologies** and skilled in using version control systems.
* Experience in creating **High Level Design and Detailed Design** in the Design phase.

# TECHNICAL SKILLS

* Hadoop Technologies : Amazon EMR, Cloudera Hadoop Distribution and Hortonworks.
* Hadoop Ecosystem : HDFS, Hive, Sqoop, Spark, Spark-SQL, Spark Streaming.
* Version Control : GIT
* NoSQL Databases : HBase, MongoDB
* Programming : C, C++, Java, Scala, Python, VC++, SQL,MAVEN
* RDBMS : MySQL, SQL Server, Redshift, Snowflake
* IDE : Eclipse, Microsoft Visual Studio
* Operating System : Linux (RHEL, CentOS), Windows 7onwards
* ClusterManagementTools : Hortonworks, Cloudera, EMR and AZURE

# CERTIFICATIONS

* AWS Certified Developer - Associate
* SAFE5 Certified Practitioner

# EDUCATION

**Bachelors in Computer Applications, JNTU University, Hyderabad, India**  **June 2009**

# PROJECT DETAILS

**Client:** Vanguard| **Role**: SR Data Engineer

**Duration:** May-2023- Till Date

**Responsibilities:**

* Proficient in developing big data applications using **PySpark** for distributed data processing and analysis.
* Processing with java, spring boot and spring webflux
* Experienced in leveraging **AWS Glue** for ETL (Extract, Transform, Load) tasks, including data integration, transformation, and orchestration.
* Designed and implemented data pipelines using PySpark and AWS Glue to ingest, cleanse, transform, and load large-scale datasets from various sources into AWS data lakes and data warehouses.
* Designed and implemented data processing pipelines using Apache Spark and Scala, resulting in a [X%] improvement in processing efficiency.
* Implemented real-time data streaming solutions using Apache Kafka for ingesting and processing streaming dat
* Demonstrated expertise in optimizing PySpark jobs for performance and scalability, including partitioning, caching, and leveraging Spark SQL optimizations.
* Hands-on experience in deploying and managing PySpark applications on AWS cloud infrastructure, utilizing services like **Amazon EMR** (Elastic MapReduce) for scalable and cost-effective big data processing.
* Proficient in working with AWS services such as Amazon **S3**, Amazon Redshift, Amazon RDS, and **AWS Lambda** for data storage, querying, and serverless computing.
* Skilled in integrating PySpark applications with other AWS services and third-party tools for monitoring, logging, and automation, using services like **AWS CloudWatch, AWS Step Functions.**
* Familiarity with Python programming language and its ecosystem, including libraries like **pandas, NumPy** for data manipulation, analysis.
* Demonstrated expertise in integrating **Apache Iceberg** with big data processing frameworks like Apache Spark and Presto for querying and analyzing large datasets with high performance and reliability.
* Strong understanding of data lake architecture principles and best practices for organizing and governing data using Apache Iceberg as a foundational layer.
* Familiarity with **Jenkins** plugins and integrations for Big Data development and operations tasks, including version control systems (e.g., **Git**), build tools (e.g., **Maven**)

**Technologies / Environment:** Hadoop, Hive, Spark, Python, Scala, DB2, SQL Server, Oracle, java, HBase, MongoDB, Kafka, AWS Athena, S3, AWS Glue

**Client:** Infosys| **Role**: Senior Consultant

**Duration:** Feb-2022- Feb-2023

**Responsibilities:**

* Efficiently Ingested Data from RDBMS sources to hive using **spark**.
* Developed and implemented ETL (Extract, Transform and Load) processes using **PySpark** and **AWS Glue** to ingest and process large volumes of data efficiently.
* Successfully transforming the data to subsequent layers on applying Business logic.
* Developed Framework in **Spark** to ingest tables from **RDBMS** by configuring the Metadata information. This includes Audit and reconciliation stats.
* Used **Python** and **Scala** (Functional Programming) to develop the framework in Spark.
* Did efficient performance tuning of Spark batch jobs.
* Created pipelines to stage data extracts for Tableau consumption.
* Written and read data in **Parquet** file format.
* Conducted code reviews and provided mentorship to junior team members on Scala best practices and Big Data technologies.
* Consumed and written data to NOSQL’s like **MongoDb** and **HBase**.
* Developed Real-time streaming applications with **Kafka** as messaging system.
* Applied multiple and complex data transformation in Spark using **Dataframes**.
* Operate as an Individual contributor by providing architecture, design and development of new business capabilities and ensuring all aspects of the solution follow accepted best practices, meet salability and resiliency expectations, and will be efficiently supported.
* Designing big data implementation procedures to ensure they comply with internal and external regulations.
* Collaborated with cross-functional teams to gather requirements and design scalable data architectures.
* Batch processing with java/spark SQL
* Have scheduled jobs in **Apache Airflow**

**Client:** Verizon| **Role**: Engineer III Consultant

**Duration:** Mar-2018- Jan-2022

**Responsibilities:**

* Successfully Ingested Data from source to Hive using **sqoop**.
* Successfully transformed the data to subsequent layers on applying Business logic.
* Scheduling Jobs via **OOZIE**, monitoring and enhancing the performance.
* Created data pipeline for few sources using **Streamsets**.
* Used java script for the front-end maintenance and new function implementation
* Created reconciliation scripts for matching the count of ingested data to source.
* Consuming complex **JSON** and **XML** data and creating structure using Hadoop in the form of Hive tables on **S3**.
* Developed Unique Framework in **Spark** to ingest tables from **RDBMS** which is a backend for Guidewire by configuring the Metadata information. This includes Audit and reconciliation stats.
* Used **Scala** (Functional Programming) to develop the framework in Spark.
* Ingested data from **SQLServer** and **Netezza** Sources.
* Improved Performance tuning of Spark batch jobs.
* Written and read data in **Parquet** and **ORC** file formats.
* Applied data transformation in Spark using **Dataframes**.
* Created layer out of **AWS Athena** for business querying purpose.
* Troubleshooting job failures by reviewing the log files.
* Analyst, transform and validate files in streaming with java8 / spark
* Development of result WS web service using java and spring boot.
* Ingested data from **SQLServer**, **Oracle, DB2** Sources to **IBM Sailfish**.

**Technologies / Environment:** Hadoop, Hive, Spark, Scala, ControlM, Nifi, IBM Sailfish, DB2, SQL Server, Oracle, HBase, MongoDB, Kafka, Cloudera

**Client:** Verizon| **Role**: Analyst System Dev

**Duration:** Jun-2013- Mar-2018

**Responsibilities:**

* Involved in modifying the product as per customer requirements.
* Created **complex SQL Queries** which are used in code for reporting purposes
* Actively Participated in Migrating the code from **DB2 to Oracle** with Rac Architecture.
* Written code in changing the WF related changes for all the new requirements.

**Technologies / Environment:** DB2, Oracle, C, C++, JAVA, Shell, Perl, Eclipse, LINUX, SOLARIS.

**Client:** RSA Securities| **Role**: Software Developer

**Duration:** Jan-2011- Jun-2013

**Responsibilities:**

* Involved in modifying the product as per customer requirements.
* SSLC library upgrade and bug fixing in the upgrade.
* Involved in developing the new CMPv2 server compliant to latest rfc(rfc-4210).

**Technologies / Environment:** C, C++, VC++, Shell, Microsoft Visual Studio.

**Client:** Rapiscan| **Role**: Software Consultant

**Duration:**Jun-2009- Dec-2010

**Responsibilities:**

* Involved in developing User Interactive GUI.
* Interfaced with hardware in order to send and receive signals.
* Generated the reports and also displayed the data in the form of graphs.
* Manual testing of the Functionality.

**Technologies / Environment:** VC++, MFC, Oracle 10g.