**Kranthi Kumar**

**Phone: 469-933-3805**

**Email: kranthikumar12345@gmail.com**

**SUMMARY:**

* Over **10+ years** of strong experience as a **Data Engineer** in data mining large data sets of Structured, Data Acquisition, Data Validation, Predictive modeling, Data modeling, Data Visualization, Web Crawling, and Web Scraping.
* Adept in programming languages like **Python, Scala, Apache Spark, PySpark,** including **Big Data technologies like Hadoop, Hive,** **HBase, Kafka, Zookeeper, Ganglia UI.**
* Hands on experience in installation, configuration, and supporting **Clusters using Microsoft Azure Databricks** and **CloudFormation.**
* Skilled in **Microsoft** **Azure Big Data Processing** (ETL) and analytics using **Azure Data Bricks and Azure Data Factory.**
* In-depth understanding in **Spark Core - Spark SQL, Data Frames, Spark Streaming, MLlib, GraphX**.
* Hands on Experience in designing and developing **Spark applications using PySpark**.
* Expertise in **Spark RDD transformations, actions, Data Frame transformations,** case classes for the required input data and performed data transformations using Spark-Core and exp in writing Spark applications using **Python and Scala.**
* Good understanding of **NoSQL databases** like **HBase, Cassandra, and MongoDB**.
* Strong Experience in working with Databases like **SnowFlake, MySQL,** and proficiency in writing **complex SQL queries**.
* Experienced working with **JIRA** for project management, **GIT** for source code management.
* Experience in creating **PowerShell scripts** loading data from SnowFlake in the production env.
* Solid experience in working with **csv, text, sequential, Avro, parquet, orc, and Json** formats of data.
* Solid experience in working with **Delta Tables/ Delta Lake in Azure Ecosystem.**
* Used **Pandas, NumPy, SciPy, Matplotlib, Scikit-learn, Spark, Snowflake, NLTK** in Python for developing various **Machine Learning** algorithms and utilized machine learning algorithms such as **Linear regression, Random Forests** for data analysis.
* Experience in scheduling jobs using **Azure Data Factory and Azure DataBricks**.

**TECHNICAL SKILLS:**

**Big Data Ecosystem:** Azure DataFactory, Azure Databricks, ADLS Gen2, Cluster Management, Microsoft Azure Portal, Spark, Kafka, STORM, Hive, HBase, Zookeeper, Snowflake.

**Tools and Utilities:** Tableau 8/10, JIRA, GIT, TIBCO Spotfire, POWER BI, ADLCopy, Azure Storage Explorer, DBFS Explorer.

**Languages:**Python, Scala, SQL, HTML, JavaScript, Unix Shell Script.

**Databases:**SQL Server, HBASE, PostgresSQL, Azure SQL.

**Operating Systems:**Microsoft Windows 8/7/XP, Linux and UNIX

**PROFESSIONAL EXPERIENCE:**

**Wells Fargo Bank,** **Dallas, TX Oct 2018-Present**

**Sr. Data Engineering Team Lead**

* Good understanding and hands on experience on **Azure Ecosystem** and services like **Azure storage (Blob, table), Azure Logic Apps, Azure Service Bus, Azure Key Vault, Cosmos DB, Azure SQL, Azure Data factory, Azure Databricks, Polybase, Azure Datalake, Azure Active Directory.**
* Developed **Spark applications using Scala and Spark-SQL** for data extraction, transformation, and aggregation from multiple data file formats to uncover insights into the customer usage patterns.
* Experience in **performance tuning of Spark Applications** for setting right Batch Interval time, correct level of **Parallelism and memory tuning**.
* Used **Zeppelin & Jupyter** notebooks, and **Spark-Shell** to develop, test, and analyze Spark jobs before scheduling customized Spark jobs.
* Experienced in migrating on premises applications to **Microsoft Azure.**
* Hands on experience in creating CI/CD pipelines using **Azure DevOps.**
* Created pipelines to move data from **Azure Blob storage thru Azure Data Warehouse using ADF.**
* Created tools using C#.Net to read Azure blob/ file information in Azure Storage.
* Enhanced Metadata design for the pipelines.
* Developed Sales History pipelines to process files from **Azure Blob to Azure Data Warehouse using polybase**.
* Created External tables to load staging data using a poly base for the data stored in Azure storage.
* Used external tables to load data from other data sources **(Azure Blob, Azure Data Lake HDFS).**
* Imported on-premises data to Azure environment using **Azure Data Factory**
* Developed SSIS packages and deployed in Azure environment.
* Implemented Azure Application Insights to store user activities and error logging
* Created **delta and full load pipelines** in Azure Data Factory
* Created on-demand tables on notebook using **Python and PySpark.**
* Optimized existing algorithms in Hadoop using **Spark Context, Spark-SQL, Data Frames and Pair RDD's**.
* Developed pipeline for POC to compare performance/efficiency while running pipeline using the **Databricks Spark cluster~~.~~**
* Responsible for loading customer's data and event logs from **Kafka into HBase using REST API.**
* Developed **ETL pipelines** in and out of data warehouse using combination of Python and Snowflake’s **SnowSQL**
* Wrote **UDF’s in Scala** and Store procedures to meet specific business requirements.
* Participated in **code reviews** with peers to ensure proper test coverage and consistent code standards.
* Created **DDL's** for tables and executed them to create tables in the warehouse for **ETL data loads**.
* Led **Process change** within the Data Engineering team to **enhance productivity and quality** of workflows.
* Performed **cross-team and internal trainings** for multiple deliverables being produced by data engineering team.
* **Handled escalations and resource assignments** **as a Team Lead,** for several tasks/tickets in production env.

**Environment:** Hadoop, Azure Data Factory, Azure Storage, Azure SQL Server, Azure Data Warehouse, Power BI, Azure Key Vault, Azure Data Bricks, PowerShell, Azure DevOps, Git, Azure Active Directory, Spark, Spark MLlib, Kafka, HBase, Scala, Python, Tableau, SQL.

**Overhead Door Corporation, Dallas, TX Oct 2017- Oct 2018**

**Sr. Data Engineer**

* Design and implementation of **Big Data Analytics** architecture using **MapReduce Programming, Spark, Hive, HBase, Kafka, Zookeeper, Ganglia UI.**
* Utilize tools like **Hive, Spark** to leverage insights from **Client CRM and log files**.
* Worked on **Spark SQL** in creating data frames by loading data from Hive tables.
* Developed batch scripts to fetch data from **Database** and performed transformations in **Databricks using PySpark.**
* Used **Spark Streaming APIs** to perform transformations and actions to build common learner data model to get data from **Kafka** real-time to **HBase**.
* Used **Apache Kafka and Spark Streaming** to get the data from live stream rest API connections.
* Ensured data integrity & quality to fast-track smooth data onboarding and product delivery to clients using **Python, bash scripting, Hive** etc.
* Led the **code standardization efforts across the Data Engineering team** to ensure common code structure deployment across various use cases and clients.
* Led **QC initiative** across company for Data Engineering team, including various stages and levels of validation.
* Implement and facilitate **PySpark transition** for existing infrastructure within the Data Engineering group.
* Utilize **Databricks Notebook and Jupyter notebooks to create PySpark enabled notebooks** which perform data wrangling and data manipulation operations.

**Environmen**t: Azure, SparkSQL, PySpark, Hive, Python, Scala, MySQL

**Ryder System Inc. Aug 2011 - Oct 2017**

**Big Data Engineer/Developer**

* Responsible for **Big Data** **analysis, develop, and design solutions for ETL and Business Intelligence platforms.**
* Installed **Kafka** on **Hadoop** cluster and configured producer and consumer in Java to establish connection from source to HDFS with popular hash tags.
* Load real time data from various data sources into **HDFS** using **Kafka**.
* Worked on reading multiple data formats on **HDFS** using **Python**.
* Implemented **Spark** using **Python (PySpark)** and **SparkSQL** for faster testing and processing of data.
* Involved in converting **Hive/SQL** queries into Spark transformations using API’s like **Spark SQL, Data Frames,** and **Python**.
* Analyzed the **SQL scripts** and designed the solution to implement using **Python.**
* Tuning the Spark by improving the performance and optimization of the existing algorithms in Hadoop using **Spark Context, SparkSQL, Data Frame, Pair RDD, & Spark YARN.**
* Used various concepts in Spark like **broadcast variables, caching, dynamic allocation,** etc. to design more scalable spark applications.
* Involved in converting **Map Reduce** programs into Spark transformations using **Spark RDD on Python.**
* Worked with **NoSQL** databases like **HBase** in creating tables to load large sets of data from various sources.
* Design and develop the **HBase** target schema.
* Worked on visualizing the reports using **Tableau**.
* Experience in **Commissioning and Decommissioning nodes** in the cluster.
* Experience in setting up **Hadoop quotas**, and configuration and maintenance of **YARN schedulers.**
* **Support DevOps** teams with transition for increased security and data privacy policies.
* **Led code reviews across Hadoop Engineering teams** to QC results and ensure quality standards are adhered.
* **Designed ETL processes using Informatica** to load data from Flat Files, Oracle, and Excel files to target Oracle Data Warehouse database.

**Environment:** Hadoop, Map Reduce, Spark, Spark MLlib, Kafka, HBase, HIVE, Scala, Python, SQL, HTML, Tableau.