**Sr. Data Engineer**

**Name: Ravina**

**Phone:** **+1 (206) 251-7999**

**Email:** **k15.ravina@hotmail.com**

**LinkedIn:** **https://www.linkedin.com/in/ravina-k-61a8422bb/**

**SUMMARY**

* Overall, 10 years of experience as Data Engineer, Python Developer. Proficient in designing, documenting, development, and implementation of data models for enterprise - level applications. Background in Data Lake, Data Warehousing, ETL Data pipeline & Data Visualization. Proficient in Big data storage, processing, analysis, and reporting on all major Cloud vendors-AWS, Azure, GCP.
* Functioned as Data Engineer responsible for data modelling, data migration, design, preparing ETL pipelines for both cloud and on Exadata.
* 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.
* Good knowledge on Apache Spark components including SPARK CORE, SPARK SQL, SPARK STREAMING and SPARK MLLIB.
* Experience exclusively on Big Data Ecosystem using HADOOP framework and related technologies such as HDFS, MapReduce HIVE, PIG, HBASE, STORM, YARN, OOZIE,
* SQOOP, Airflow and Zookeeper and includes working experience in Spark Core, Spark SQL, Spark Streaming, Scala, and Kafka.
* Hands-on experience with Amazon EC2, S3, RDS, IAM, Auto Scaling, CloudWatch, SNS, Athena, Glue, Kinesis, Lambda, EMR, Redshift, DynamoDB and other services of the AWS family.
* Extensively worked on Spark using Scala on cluster for computational (analytics), installed it on top of Hadoop performed advanced analytical application by making use of Spark with Hive and SQL/PostgreSQL/Snowflake.
* Utilized Kubernetes and Docker for the runtime environment for the CI/CD system to build, test, and deploy.
* Migrated an existing on-premises application to AWS. Used AWS services like EC2 and S3 for small data sets processing and storage, experienced in maintaining the Hadoop cluster on AWS EMR. Hands-on experience with Amazon RDS, Auto Scaling, CloudWatch, SNS, Athena, Glue, Kinesis, Redshift, DynamoDB and other services of the AWS family.
* Involved in all phases of ETL life cycle from scope analysis, design, and build through production support.
* Designed and developed logical and physical data models that utilize concept such as Star Schema, Snowflake Schema and Slowly Changing Dimensions.
* Design and develop Spark applications using PySpark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage pattern.
* Experienced in Data Modeling &Data Analysis experience using Dimensional Data Modeling and Relational Data Modeling, Star Schema/Snowflake Modeling, FACT & Dimensions tables, Physical &Logical Data Modeling.
* Developed and maintained ETL pipelines using IBM Datastage/Matillion to extract data from various sources such as databases, APIs, and flat files, transform it according to business logic, and load it into target data warehouses or data lakes.
* Expertise in Python data extraction and data manipulation, and widely used python libraries like NumPy, Pandas, and Matplotlib for data analysis.
* Extensively worked with Teradata utilities Fast export and Multi Load to export and load data to/from different source systems including flat files.
* Implemented data quality checks and governance processes within ETL pipelines to ensure the accuracy, consistency, and completeness of data, adhering to established data quality standards and best practices.
* Extensively worked on other machine learning libraries such as Seaborn, Scikit learn for machine learning and familiar working with TensorFlow, NLTK for deep learning
* Actively involved in each phase of software Development life cycle (SDLC) and experience in Agile Software Methodology.
* Knowledgeable about scaling Neo4j deployments horizontally and vertically to accommodate growing data volumes and ensuring high availability through replication and clustering
* Excellent understanding of best practices of Enterprise Data Warehouse and involved in Full life cycle development of Data Warehousing
* Expertise in using Sqoop & Spark to load data from MySQL/Oracle to HDFS or HBase.
* Well versed in using ETL methodology for supporting corporate-wide- solution using Informatica.
* Responsible for writing Unit Tests and deploying production level code through the help of Git version control.
* Experience in Web Application Development, hands on experience of Python/Django, Java/Spring, HTML5/CSS3, Bootstrap, JavaScript, and Angular, React, jQuery and JSON/AJAX.
* Experience in working on different operating systems Windows, Linux, UNIX, and Ubuntu.

**TECHNICAL SKILLS**

**ETL Tools:**AWS Glue, Azure Data Factory, Airflow, Spark, Sqoop, Flume, Apache Kafka, Spark Streaming

**NoSQL Databases:**MongoDB, Cassandra, Amazon DynamoDB, HBase

**Data Warehouse:**AWS RedShift, Google Cloud Storage, Snowflake, Teradata, Azure Synapse

**SQL Databases:**Oracle DB, Microsoft SQL Server, IBM DB2, PostgreSQL, Teradata, Azure SQL Database, Amazon RDS, GCP Cloud SQL, GCP Cloud Spanner

**Web Development:**HTML, XML, JSON, CSS, JQUERY, JavaScript

**Monitoring Tools:**Splunk, Chef, Nagios, ELK

**Source Code Management:**J Frog Artifactory, Nexus, GitHub, Code Commit

**Containerization:**Docker & Docker Hub, Kubernetes, OpenShift

**Hadoop Distribution:**Cloudera, Hortonworks, MapR, AWS EMR, Azure HDInsight, GCP DataProc

**Programming and Scripting:**Spark Scala, Python, Java, MySQL, PostgreSQL, Shell Scripting, Pig, HiveQL

**AWS:**EC2, S3, Glacier, Redshift, RDS, EMR, Lambda, Glue, CloudWatch, Kinesis, CloudFront, Route53, DynamoDB, Code Pipeline, EKS, Athena, Quick Sight

**Hadoop Tools:** HDFS, HBase, Hive, YARN, MapReduce, Pig, HIVE, Apache Storm, Sqoop, Oozie, Zookeeper, Spark, SOLR, Atlas

**Build & Development Tools:**Jenkins, Maven, Gradle, Bamboo

**Methodologies:**Agile/Scrum, Waterfall

**PROFESSIONAL EXPERIENCE**

**Client: Wintrust Financial Corporation Rosemont, IL Oct 2023 – Till Date**

**Role: AWS DATA ENGINEER**

**Responsibilities:**

* Setting Infrastructure as a code (Iaas) in the AWS cloud platform from the scratch through CloudFormation templates by configuring and integrating appropriate AWS services as per the business requirement
* Deployed Snowflake following best practices, and provide subject matter expertise in data warehousing, specifically with Snowflake.
* Developed the PySpark code for AWS Glue jobs and for EMR.
* Implemented MDM solutions using IBM Infosphere MDM or similar platforms to create a single, authoritative source of master data across the enterprise, enabling consistent and reliable business insights.
* Designing the business requirement collection approach based on the project scope and SDLC methodology.
* Led the design, development, and deployment of MDM solutions using platforms such as Informatica MDM, IBM Infosphere MDM, or Talend MDM, establishing a single, authoritative source of master data across the enterprise.
* Facilitated effective communication between data engineering, data science, and business teams to ensure alignment on project goals and objectives.
* Proficient in conveying complex technical concepts to non-technical stakeholders through clear and concise communication
* Implemented security measures in Neo4j, including authentication, authorization, and encryption, to safeguard sensitive data and comply with regulatory requirements.
* Adapted ETL processes for cloud data platforms such as AWS Redshift, Google BigQuery, or Snowflake, leveraging native integration features and optimizing data transfer and processing for cloud environments.
* Involved in converting Hive/SQL queries into Spark Transformations using Spark RDDs and Scala and involved in using SQOOP for importing and exporting data between RDBMS and HDFS.
* Experience in building models with deep learning frameworks like TensorFlow, PyTorch, and Keras.
* Experienced in Python data manipulation for loading and extraction as well as with python libraries such as matplotlib, NumPy, SciPy and Pandas for data analysis.
* Created monitors, alarms, notifications and logs for Lambda functions, Glue Jobs, EC2 hosts using CloudWatch and used AWS Glue for the data transformation, validate and data cleansing.
* Worked with cloud-based technology like Redshift, S3, AWS, EC2 Machine, etc. and extracting the data from the Oracle financials and the Redshift database and Create Glue jobs in AWS and load incremental data to S3 staging area and persistence area.
* Handled importing data from various data sources, performed transformations using Hive, Map Reduce, and loaded data into HDFS.
* Experience in working with Dataflow pipelines using Apache Beam API to migrate the data from various sources to Big Query.
* Deployed applications using Jenkins framework integrating Git- version control with it.
* Used the Agile Scrum methodology to build the different phases of Software development life cycle.
* Improving the performance and optimization of the existing algorithms in Hadoop using Spark Context, Spark-SQL, Data Frame, Pair RDD's, YARN.
* Experience in working with Dataflow pipelines using Apache Beam API to migrate the data from various sources to Big Query.
* Scheduled Airflow DAGs to run multiple Hive and Pig jobs, which independently run with time and data availability and Performed Exploratory Data Analysis and Data Visualizations using Python, and Tableau
* Designed, developed, and implemented ETL pipelines using python API (PySpark) of Apache Spark on AWS EMR.
* Proficient in writing unit tests using testing frameworks such as JUnit, PyTest, or NUnit to ensure the reliability and accuracy of code.
* Developed Spark applications using Pyspark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats.
* Worked on Apache Dataflow to migrate the data from SQL server, ERP systems and FTP to Big Query in GCP.
* Implemented various data structures such as arrays, linked lists, trees, and graphs to support data manipulation and analysis tasks.
* Implemented functional testing strategies to validate the correctness and functionality of data pipelines and analytical algorithms.
* Used SSIS to build automated multi-dimensional cubes.
* Used Spark Streaming to receive real time data from the Kafka and store the stream data to HDFS using Python and NoSQL databases such as HBase and Cassandra.
* Worked on Dimensional and Relational Data Modelling using Star and Snowflake Schemas, OLTP/OLAP system, Conceptual, Logical and Physical data modelling.
* Automated the data processing with Oozie to automate data loading into the Hadoop Distributed File System.
* Developed Automation Regressing Scripts for validation of ETL process between multiple databases like AWS Redshift, Oracle, MongoDB, T-SQL, and SQL Server using Python.
* Create Data pipelines for Kafka cluster and process the data by using spark streaming and worked on streaming data to consume data from KAFKA topics and load the data to landing area for reporting in near real time.
* Used Informatica power center to Extract, Transform and Load data into Netezza Data Warehouse from various sources like Oracle and flat files.
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDDs, Python and Scala.
* Good knowledge about configuration management tools like Bitbucket/GitHub and Bamboo (CICD).
* Collect the data using Spark streaming and dump into HBase and Cassandra. Used the Spark- Cassandra Connector to load data to and from Cassandra.
* Hands-on experience with designing, deploying and maintenance of MDM systems like AirWatch, Workspace.

**Client: Couth InfoTech Pvt. Ltd Jan 2020 - Feb 2023**

**Role: DATA ENGINEER**

**Responsibilities:**

* Experience in building and architecting multiple Data pipelines, end to end ETL and ELT process for Data ingestion and transformation in GCP.
* Developed various Mappings with the collection of all Sources, Targets, and Transformations using Informatica Designed.
* Integrated unit testing into continuous integration and deployment pipelines to automate test execution and maintain code quality.
* Executed performance tests to assess the scalability and efficiency of data processing systems and algorithms.
* Demonstrated expertise in designing and implementing efficient data structures to optimize data processing and storage.
* Worked on partitioning and clustering of datasets in Big Query
* Created Dataflow pipelines using Apache Beam API in Java.
* Implemented solutions for ingesting data from various sources and processing the Data-at-Rest using Big Data technologies such as Hadoop, Map Reduce Frameworks, HBase, and Hive.
* Experience in Amazon AWS services such as EMR, EC2, S3, CloudFormation, RedShift which provides fast and efficient processing of Big Data.
* 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.
* Written ETL jobs to visualize the data and generate reports from MySQL database using DataStage.
* Involved in all the steps and scope of the project reference data approach to MDM, have created a Data Dictionary and Mapping from Sources to the Target in MDM Data Model.
* Responsible for developing data pipeline using Spark, Scala, Apache Kafka to ingest the data from CSL source and store in HDFS protected folder.
* Developed Oozie workflow engine to run multiple Hive, Pig, Tealeaf, Mongo DB, Git, Sqoop and Spark jobs.
* Designed graph data models using Neo4j, leveraging its graph-based structure to represent complex relationships between entities in data.
* Integrated Neo4j with various data sources and ETL pipelines, enabling seamless data ingestion, transformation, and loading processes into the graph database.
* Developed custom Kafka producer and consumer for different publishing and subscribing to Kafka topics.
* Experienced in Python data manipulation for loading and extraction as well as with python libraries such as matplotlib, NumPy, SciPy and Pandas for data analysis.
* Design and implement multiple ETL solutions with various data sources by extensive SQL Scripting, ETL tools, Python, Shell Scripting, and scheduling tools. Data profiling and data wrangling of XML, Web feeds and file handling using python, Unix and SQL.
* Involved in converting Hive/SQL queries into Spark Transformations using Spark RDDs and Scala and involved in using SQOOP for importing and exporting data between RDBMS and HDFS
* Created continuous integration and continuous delivery (CI/CD) pipeline on AWS that helps to automate steps in software delivery process.
* Involved in building Data Models and Dimensional Modeling with 3NF, Star and Snowflake schemas for OLAP and Operational data store (ODS) applications.
* Developed Producer API and Consumer API to publish and subscribe to stream of events in one or more topics.
* Developed and deployed data pipeline in cloud such as AWS and GCP
* Implemented Lambda to configure Dynamo DB Autoscaling feature and implemented Data Access Layer to access AWS DynamoDB data.
* Involved in Relational and Dimensional Data modeling for creating Logical and Physical Design of Database and ER Diagrams with all related entities and relationship with each entity based on the rules provided by the business manager using ER Studio.
* Created Data Lake by extracting customer's data from various data sources into this includes data from Teradata, Mainframes, RDBMS, CSV, and Excel.
* Worked on confluence and Jira was skilled in data visualization like Matplotlib and seaborn library.
* Hands of experience in GCP, Big Query, GCS bucket, G - cloud function, cloud dataflow, Pub/sub cloud shell, GSUTIL, BQ command line utilities, Data Proc, Stack driver.

**Client: EAIESB Software Solutions Oct 2017 - Dec 2019**

**Role: AZURE DATA ENGINEER**

**Responsibilities:**

* Extracted, Transformed and Loaded data from Sources Systems to Azure Data Storage services using a combination of Azure Data Factory, Spark SQL, and U-SQL Azure Data Lake Analytics.
* Data is Ingested to one or more Azure Services- (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Worked on Azure Services like IaaS, PaaS and worked on storage like Blob (Page and Block), SQL Azure.
* Implemented OLAP multi-dimensional functionality using Azure SQL Data Warehouse.
* Retrieved data using Azure SQL and Azure ML which is used to build, test, and predict the data.
* Worked on Cloud databases such as Azure SQL Database, SQL managed instance, SQL Elastic pool on Azure, and SQL server.
* Designing and implementing performant data ingestion pipelines from multiple sources using Apache Spark and/or Azure Databricks
* Architect & implement medium to large scale BI solutions on Azure using Azure Data Platform services (Azure Data Lake, Data Factory, Data Lake Analytics, Stream Analytics, Azure SQL DW, HDInsight/Databricks, NoSQL DB)
* Involved in Migrating Objects from Teradata to Snowflake and created Snow pipe for continuous data load.
* Created continuous integration and continuous delivery (CI/CD) pipeline on AZURE that helps to automate steps in the software delivery process.
* Data ingestion to one or more Azure services (Azure Data Lake, Azure Storage, AZURE SQL, Azure synapse) and processing the data in Azure Databricks
* Involved in converting Hive/SQL queries into Spark transformations using Spark RDD's and PySpark.
* Handled importing of data from various data sources, performed transformations using Hive, and loaded data into HDFS.
* Developed PySpark pipelines which transform the raw data from several formats to parquet files for consumption by downstream systems.
* Builds and maintains complex statistical routines using PCSAS macros, Enterprise Guide, PL/SQL, and software written by myself and others.
* Developed ETL solutions using Spark SQL in Azure Databricks for data extraction, transformation and aggregation from multiple file formats and data sources for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Created dashboards in Tableau that were published to the internal team for review and further data analysis and customization using filters and actions.
* Develop Reports and Dashboards in Splunk. Utilized Splunk Machine Learning Toolkit for modeling Clustering model to detection log.
* 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.
* Performed Data Analysis, Data Migration, Data Cleansing, Transformation, Integration, Data Import, and Data Export through Python
* Experienced in running queries using Impala and used BI tools and reporting tool (tableau) to run ad-hoc queries directly on Hadoop.
* Hands on experience on fetching the live stream data from UDB into HBase table using PySpark streaming and Apache Kafka.

**Client: L&T Jul 2014 - Sep 2017**

**Role: Data Engineer**

**Responsibilities:**

* Involved in requirements gathering and analysis from the existing system. Captured requirements using Use Cases and Sequence Diagrams.
* Designed web portals using HTML & used Java script, AngularJS, AJAX.
* Used Spring IOC for dependency injection and Spring AOP for cross cutting concerns like logging, security, transaction management.
* Integrated Spring JDBC for the persistence layer
* Developed code in annotation driven Spring IoC and Core Java (extensive use of Collection framework and Multithreading using Executor Framework, Callable, and Future).
* Developed DAO Classes and written SQL for accessing Data from the database.
* Used XML for data exchange and developed Web Services.
* Deployment of the application into WebSphere Application Server.
* Implemented ANT build tools to build jar and war files and deployed war files to target servers.
* Implemented test cases with JUnit.
* Used RAD for developing and debugging the application.
* Utilized Rational Clear Case as a version control system and for code management.
* Coordinated with the QA team and participated in testing.
* Involved in Bug Fixing of the application.