BINITA THAPA

Irving, Tx • 469 708 6975 • binitathapa@621gmail.com •LinkedIn

Professional Summary

- Over **5+ years** of competitive work experience in Design, Development and Maintenance of **Data Engineering applications (Data Warehouse/ETL/Hadoop)**.
- Experience in **Software Development Life Cycle (SDLC)** including Requirements Analysis, Design Specification and Testing as per Cycle in both **Waterfall and Agile** methodologies.
- Experience in designing and developing applications in **Spark** using **Scala and PySpark** to compare the performance of **Spark with Hive and SQL/Oracle.**
- Experience in implementing applications on **Spark** frameworks using **Scala.**
- Experience in using **Python** for **Data Engineering** and **Modeling**
- Experience in developing customized **UDF's** in **Python** to extend **Hive** functionality.
- Experienced in developing Web Services with **Python** programming language.
- Experience in designing **star schema, Snowflake schema** for Data Warehouse, ODS architecture.
- Experience with Informatica **(ETL Tool)** for Data Extraction, Transformation and Loading.
- Scheduled reports for Daily, Weekly, and Monthly reports on sales and marketing information for various categories and regions based on business needs using the **Power BI** reporting tool.
- Proficiency in designing scalable and efficient data architectures on Azure, leveraging services like **Azure Data Lake, Azure Data Factory, Azure Data Bricks, Azure Synapse DB**, and **Power BI.**
- Experience working with **Snowflake** Multi cluster and virtual warehouses in **Snowflake.**
- Good at Manage hosting plans for **Azure Infrastructure**, implementing & deploying workloads on **Azure virtual machines (VMs).**
- Experience in working with **AWS** cloud services **(VPC, EC2, S3, Redshift, Data Pipeline, EMR, DynamoDB, Lambda and SQS).**
- Experienced with **AWS** batch processing of data sources using **Apache Spark.**
- Experience in writing **Map Reduce** jobs to perform data cleansing and preprocessing.
- Experience in building **PySpark** and **Spark-Scala** applications for interactive analysis, batch processing, and stream processing.
- Experience in data analysis using **HIVE, HBase and custom Map Reduce** programs.
- Experience in creating shell scripts to push data loads from various sources from the edge nodes onto the **HDFS.**
- Experienced in working with different data formats **CSV, JSON.**
- Experience in working with Hive data warehouse tool-creating tables, data distribution by implementing partitioning and bucketing, writing, and optimizing the **HiveQL queries.**
- Hands on experience in **SQL and NOSQL** database such as **Snowflake, HBase, Cassandra and MongoDB.**
- Involved in all phases of software development life cycle in **Agile, Scrum and Waterfall** management process.
- A self-motivated exuberant learner and adequate with challenging projects and work in ambiguity to solve complex problems independently or in the collaborative team.
- Strong skills in analytical, presentation, communication, problem solving with the ability to work independently as well as in a team and had the ability to follow the best practices and principles defined for the team.

Technical Skills:

Databases	Snowflake, AWS RDS, Teradata, Oracle, MySQL, Microsoft SQL, PostgreSQL.
NoSQL Databases	MongoDB, Hadoop HBase, and Apache Cassandra.
Programming Languages	Python, SQL, Scala, MATLAB.
Cloud Technologies	AWS, Docker
Data Formats	CSV, JSON
Querying Languages	SQL, NoSQL, PostgreSQL, MySQL, Microsoft SQL
Integration Tools	Jenkins
Scalable Data Tools	Hadoop, Hive, Apache Spark, Map Reduce, Sqoop.
Operating Systems	Red Hat Linux, Unix, Windows, macOS.
Reporting & Visualization	Tableau, Matplotlib.

Professional Experience

Client: Paychex Rochester, NY **July 2023 – Till Date**
Role: Data Engineer
Responsibilities:

- Interacted with clients to gather business and system requirements which involved documentation of processes based on the user requirements.
- Developed **Scala** based **Spark** applications for performing data cleansing, event enrichment, data aggregation, de-normalization and data preparation needed for machine learning and reporting teams to consume.
- Developed **Spark Scala** scripts for mining information and performed changes on huge datasets to handle ongoing insights and reports.
- Developed **spark** applications in **python (PySpark)** on distributed environment to load huge number of CSV files with different schema in to Hive ORC tables.
- Developed **Scala** scripts using both **Data frames/SQL/Data sets** and **RDD/MapReduce** in **Spark** for Data Aggregation, queries and writing data back into OLTP system through **Sqoop.**
- Developed story telling dashboards in **Tableau Desktop** and published them on to **Tableau Server** which allowed end users to understand the data on the fly with the usage of quick filters for on demand needed information.
- Created workflows, mappings using **Informatica ETL** and worked with different transformations such as lookup, source qualifier, update strategy, router, sequence generator, aggregator, rank, stored procedure, filter, joiner, sorter.
- Create several types of data visualizations using **Python and Tableau.**
- Designed and implemented configurable data delivery pipeline for scheduled updates to customer facing data stores built with **Python.**
- Install and Configure **Apache Airflow for S3 bucket** and snowflake data warehouse and created DAGs to run the **Airflow.**
- Built and managed data pipelines using **Azure Data Factory and Azure Data Bricks** ensuring efficient and reliable data processing and analysis workflows.
- Automated advanced **SQL queries and ETL** techniques using **Apache Airflow** to reduce boring weekly administration tasks.
- Used **Scala** to convert **Hive** / **SQL** queries into RDD transformations in **Apache Spark**.
- Used **Kafka** producer to ingest the raw data into **Kafka** topics run the **Spark Streaming** app to process clickstream events.
- Written **Hive** jobs to parse the logs and structure them in tabular format to facilitate effective querying on the log data.
- Used **GIT** to check-in and checkout code changes.
- Written multiple **MapReduce** programs for data extraction, transformation and aggregation from multiple file formats including **XML, JSON, CSV** & other compressed file formats.
- Worked with complex SQL, Stored Procedures, Triggers, and packages in large databases from various servers.
- Involved in **Agile** methodologies, daily scrum meetings, spring planning.

Environment: Spark, Scala, AWS, ETL, Kafka, Tableau, Hadoop, Python, Snowflake, HDFS, Hive, MapReduce, PySpark, Docker, Sqoop, Apache Airflow, Power BI, Teradata, JSON, MongoDB, SQL, Agile and Windows.

Client: Verizon Irving, TX **Apr 2022 - July 2023**
Role: Data Engineer
Responsibilities:

- Participated in requirement gathering session with business users and sponsors to understand and document the business requirements.
- Developed **Spark** Applications by using **Scala** and Implemented **Apache Spark** data processing project to handle data from various **RDBMS** and Streaming sources.
- Designed and implemented **Spark** jobs to support distributed data processing.
- Designed and Developed **Scala** workflows for data pull from cloud-based systems and applying transformations on it.
- Designed Data Quality Framework to perform schema validation and data profiling on **Spark (PySpark).**
- Developed various **Python** scripts to find vulnerabilities with **SQL Queries** by doing SQL injection, permission checks and analysis.

- Involved in designing optimizing **Spark SQL queries,** Data frames, import data from Data sources, perform transformations; perform read/write operations, save the results to output directory into **HDFS/AWS S3.**
- Developed **Tableau** data visualization using Cross tabs, Heat maps, Box and Whisker charts, Scatter Plots, Geographic Map, Pie Charts and Bar Charts and Density Chart.
- Prepared **ETL** design document which consists of the database structure, change data capture, Error handling, restart, and refresh strategies.
- Use **Amazon Elastic Cloud Compute** (EC2) infrastructure for computational tasks and **Simple Storage Service** (S3) as storage mechanism.
- Designed an **Apache Airflow** Data Pipeline to automate data ingestion and retrieval.
- Created pipelines in **ADF** using linked services, datasets, and pipelines to extract, transform, and load **data** from different sources like **Azure SQL,** blob storage, the **Azure SQL Data Warehouse**, the write-back tool, and backwards.
- Extract, transform, and load **data** from source systems to Azure **Data** Storage services using a combination of Azure **Data** Factory and **Data** Lake Analytics.
- **Data** ingestion to one or more Azure services (Azure **Data** Lake, Azure Storage, and Azure SQL) and processing the **data** in Azure **Data** bricks.
- Created **Sqoop** jobs with incremental load to populate Hive External tables.
- Worked on streaming pipeline that uses **Spark** to read data from **Kafka** transform it and write it to **HDFS.**
- Worked on different file formats like **Sequence files, XML files** and **Map files** using **MapReduce** Programs.
- Implemented a Continuous Delivery pipeline with **Docker** and **GitHub.**
- Responsible for modifying the code, debugging, and testing the code before deploying on the production cluster.
- Worked on designing, building, deploying, and maintaining **Mongo DB**.
- Implemented **SQL, PL/SQL** stored procedures.
- Involved in story-driven **agile** development methodology and actively participated in daily scrum meetings.

Environment: Spark, Scala, AWS, ETL, Hadoop, Python, Snowflake, HDFS, Hive, Tableau, MapReduce, PySpark, Teradata, Docker, JSON, XML, Apache Kafka, ApacheAIrflow, Power BI, SQL, PL/SQL, Agile and Windows.

Client: Western Union, Coffeyville, KS **Mar 2019 - Dec 2021**
Role: Data Engineer
Responsibilities:

- Collaborated with Business Analysts, SMEs across departments to gather business requirements, and identify workable items for further development.
- Developed various **spark** applications using **Scala** to perform various enrichment of these click stream data merged with user profile data.
- Developed highly complex **Python** and **Scala** code, which is maintainable, easy to use, and satisfies application requirements, data processing and analytics using inbuilt libraries.
- Developed **Spark code** in **Python and SparkSQL** environment for faster testing **and** processing of data and loading the data into **Spark RDD** and doing In-memory computation to generate the output response with less memory usage.
- Designed, developed, tested, and maintained **Tableau** functional reports based on user requirements.
- Developed **ETL's** in using Spark SQL, RDD, and Data Frames.
- Worked on **Scala** code base related to **Apache Spark** performing the Actions, Transformations on RDDs, Data Frames and Datasets using **SparkSQL** and **Spark** Streaming Contexts.
- Performed advanced procedures like text analytics and processing, using the in-memory computing capabilities of **Spark** using **Scala.**
- Worked on migrating **MapReduce** programs into **Spark** transformations using **Scala.**
- Worked with different feeds data like **JSON, CSV, XML** and implemented **Data Lake concept.**
- Analyzed the **SQL scripts** and designed the solution to implement using **PySpark.**
- Use **SQL** queries and other tools to perform data analysis and profiling.
- Followed agile methodology and involved in daily **SCRUM** meetings, sprint planning, showcases and retrospective.

Environment: Spark, Scala, Hadoop, Python, PySpark, AWS, MapReduce, ETL, HDFS, Hive, HBase, SQL, Agile and Windows.

Client: Verisk, Nepal **Feb 2018 - Feb 2019**
Role: Junior Data Engineer
Responsibilities:

- Involved in requirements gathering, analysis, design, development, change management, deployment.
- Developed **Spark** streaming application for **real time** sales **analytics.**
- Developed **spark** code and **spark-SQL/streaming** for faster testing and processing of data.
- Involved in converting **Map Reduce** programs into Spark transformations using **Spark RDD's** using **Scala and Python.**
- Worked with **Tableau** in analysis and creation of dashboard and user stories.
- Used **Hive** to do analysis on the data and identify different correlations.
- Created scripts to read **CSV, JSON, and parquet files** from S3 buckets in **Python** and load into **AWS S3, DynamoDB and Snowflake.**
- Designing **NoSQL** schemas in **Hbase.**
- Involved in weekly walkthroughs and inspection meetings, to verify the status of the testing efforts and the project as a whole.

Environment: Spark, Scala, Hive, JSON, AWS, MapReduce, Hadoop, Python, XML, NoSQL, HBase, and Windows.