**Professional Summary**

* **Having 12+ years** of professional experience in IT, working with various Legacy Database systems, which include work experience in Big Data technologies as well.
* Good experience on understanding of architecting, designing and operation of large scale data and analytics solutions on **Snowflake** **Cloud** **Data** **Warehouse**.
* Experience in **Requirement gathering, System analysis, handling business** and technical issues & communicating with both business and technical users.
* Hands on experience on complete **Software Development Life Cycle SDLC** for the projects using methodologies like **Agile** and hybrid methods.
* Experience in analyzing data using **Big Data** **Ecosystem** including **HDFS, Hive, HBase, Zookeeper,** **PIG, Sqoop, and Flume.**
* Knowledge and working experience on big data tools like **Hadoop**, **Azure Data Lake**, **AWS Redshift**.
* Good understanding of **Apache** **Airflow**.
* Experience in workflow scheduling with **Airflow, AWS Data Pipelines, Azure, SSIS, etc.**
* Experience on Migrating **SQL** **database** to **Azure data Lake, Azure data lake Analytics, Azure SQL** **Database, Data Bricks and Azure SQL Data warehouse** and Controlling and granting database access and Migrating On premise databases to **Azure** **Data** **lake** store using **Azure Data factory**.
* Good understanding of **Big Data Hadoop** and **Yarn architecture** along with various **Hadoop Demons** such as **Job Tracker, Task Tracker, Name Node, Data Node, Resource/Cluster Manager, and Kafka (distributed stream-processing).**
* Experience in Text Analytics, **Data Mining** solutions to various business problems and generating **data visualizations** using **SAS** and **Python**.
* Experience in Developing **Spark applications** using **Spark - SQL in Databricks** for **data** **extraction**, transformation and aggregation from multiple file formats for analyzing & transforming the data to uncover insights into the customer usage patterns.
* Good understanding of **Spark** Architecture including **Spark Core, Spark SQL, Data Frames, Spark** **Streaming, Driver Node, Worker Node, Stages, Executors** and **Tasks**.
* Strong experience and knowledge of **NoSQL** **databases** such as **MongoDB** and **Cassandra**.
* Experience in development and support knowledge on **Oracle, SQL, PL/SQL, T-SQL queries.**
* Experienced in **configuring and administering** the **Hadoop Cluster** using major **Hadoop** Distributions like **Apache Hadoop and Cloudera**.
* Solid Excellent experience in creating cloud based solutions and architecture using **Amazon Web services (Amazon EC2, Amazon S3, Amazon RDS,EMR, Glue)** and **Microsoft Azure.**
* Experienced in Technical consulting and end-to-end delivery with **architecture**, **data** **modeling**, **data** **governance** and **design** - development - implementation of solutions.
* Experience in **Big Data Hadoop Ecosystem** in ingestion, storage, querying, processing and analysis of **big data.**
* Extensive working experience in agile environment using a **CI/CD model.**
* Extensive experience working with structured data using **Spark SQL, Data frames, Hive QL,** optimizing queries, and incorporate complex **UDF's** in business logic.

**Technical Skills**

* **Big Data & Hadoop Ecosystem** : Hadoop, Hive, Solr, Apache Flume , Sqoop, Kafka , Oozie, Hue, Cloudera Manager, Stream sets
* **Cloud Technologies**: AWS, Glue,EC2, EC3,EMR, Redshift & MS Azure, Snowflake
* **Data Modeling Tools**: Erwin R9.7, ER Studio v16
* **Packages**: Microsoft Office 2019, Microsoft Project, SAP and Microsoft Visio 2019, Share point Portal Server
* **Other Tools**: VSS, SVN, CVS. Docker, CI/CD, Kubernetes
* **RDBMS / NoSQL Databases:** Oracle 19c, Teradata R15, MS SQL Server 2019, Cosmos DB,Cassandra 3.11, HBase 1.2,
* **Testing and defect tracking Tools:** HP/Mercury, Quality Center, Win Runner, MS Visio 2016 & Visual Source Safe
* **Operating System:** Windows 10/8, Unix, Sun Solaris
* **ETL/Data warehouse Tools:** Informatica 9.6, SAP Business Objects XIR3.1/XIR2, Talend, Tableau
* **Methodologies:** RAD, JAD, RUP, UML, System Development Life Cycle (SDLC), Agile, Waterfall Model.

**Professional Experience:**

**Client: Honeywell-Charlotte, NC Feb 2023- Till Date**

**Role: Lead Data Engineer**

**Responsibilities:**

* Involved in building database Model and Views utilizing **Python**, to build an interactive web based solution.
* Collaborated with other developers to handle complicated issues related with deployment of Django based applications.
* Handled development and management of front end user interfaces with help of **HTML**, **CSS**, **JQuery** and **JavaScript**.
* Modify the existing **Python/Django** modules to deliver certain format of data and adding new features.
* Automated a reporting process, using **Python**, **Luigi** (library for task workflow and dependencies), and other API’s.
* Written python scripts using **python** libraries such as **pandas**, numpy that does read/write operations on large CSV files, perform data aggregations and compare data by columns.
* Experience in integrating **Python REST API** Frame work using **Django**.
* Working experience of **Data Warehouse ETL /design** and implementation of complex big data pipelines.
* Used **Python**, **PySpark**, shell script, **oracle** scheduler, **Luigi**, **Oracle** **Pl SQL** etc.
* Analyzing the possibility of using **Kafka** for data streaming and **Python** Lite framework for file based interfaces.
* Developed **python** **scripts**, that reads customer Attributes from **AWS** **redshift** **warehouse**, process those records and uploading csv files to adobe marketing cloud.
* Developing a Marketing Cloud Service on **Amazon** **AWS**. Developed serverless application using **AWS** **Lambda**, **S3**, **Redshift, EMR** and **RDS**.
* Using Atom for coding, Postman and **SQL** Workbench for testing and debugging and **PyCharm** for developing and **Cloud** Formation authoring.
* Integrated **AWS Dynamo DB** using **AWS** **lambda** to store the values the items and backup the **Dynamo DB** streams.
* Experienced in **AWS Elastic** Beanstalk for app deployments and worked on **AWS** **lambda** with **Amazon** **kinesis**.
* Performed **S3 buckets** creation, policies and on the **IAM** role based polices and customizing the **JSON** template and **Glacier** for storage and backup on **AWS**.
* Used Jenkins to deploy our code into different environments and scheduling jobs.
* Used bug-tracking tools like Jira, confluence and version controls Git, GitLab.

**Environment**: Python, Django, Luigi, windows, Linux, MySQL, SQL, Cassandra, AWS RDS, AWS S3, AWS EC2, Kafka, JSON, Restful API, MVC architecture, GitLab, Agile, Enterprise Scheduler, Bitvise SSH Client, Scrum, JIRA, GIT.

**Client: BlackRock- NYC ,NY Aug 2021- Jan 2023**

**Role: Lead Data Engineer**

**Responsibilities:**

* Developed **data pipeline** using **Spark, Hive** and **HBase** to ingest customer behavioral data and financial histories into **Hadoop cluster** for analysis.
* Working Experience on **Azure Databricks cloud** to organize the data into notebooks and make it easy to visualize data using dashboards.
* **Performed ETL on data from different source systems to Azure Data Storage services using a** combination of Azure Data Factory, T-SQL, Spark SQL, and Azure Data Lake Analytics. Data Ingestion to one or more Azure Services - (**Azure Data Lake, Azure Storage, Azure SQL, Azure Synapse**) and processing the data in In **Azure Databricks.**
* Worked on managing the Spark Databricks by proper troubleshooting, estimation, and monitoring of the clusters.
* Performed Data Aggregation, Validation, and **Azure HD Insight** using spark scripts written in **Python.**
* Implemented **Azure Stream Analytics** for processing the real-time Geo-Spatial data for location-based targeted sales campaigns.
* Performed monitoring and management of the Hadoop cluster by using **Azure HDInsight.**
* Created partitioned tables in **Databricks** using Spark and designed a **data model** using **Azure Snowflake Datawarehouse**.
* Used **Hive**, **Impala,** and **Sqoop utilities** and **Oozie workflows** for **data extraction** and **data loading**.
* Extensively worked on **Shell scripts** for running **SAS** programs in batch mode on **UNIX**.
* Wrote **Python scripts** to parse **XML** documents and load the data in the database.
* Integrated **Nifi** with Snowflake to optimize the client session running.
* Optimized the performance of queries with modification in **TSQL queries to Snow SQL**, established joins, and created clustered indexes.
* Created stored procedures to import data into the **Elasticsearch engine**.
* Worked with **Data Governance, Data Quality,** and **Metadata** Management team to understand the project.
* Used **Spark SQL** to process a huge amount of structured datato aid in better analysis for our business teams.
* Created **HBase** tables to store various data formats of data coming from different sources.
* Responsible for importing log files from various sources into **HDFS** using **Flume**.
* Development of routines to capture and report **data quality** issues and exceptional scenarios.
* Installed and configured **Hadoop** and was responsible for maintaining cluster and managing and reviewing **Hadoop log files.**
* Involved in troubleshooting at database levels, error handling, and performance tuning of queries and procedures.
* Worked on **SAS Visual Analytics** & **SAS** Web Report Studio for data presentation and reporting.
* Extensively used **SAS/Macros** to parameterize the reports so that the user could choose the summary and sub-setting variables to be used from the web application.
* Developed **data warehouse models** in snowflake for various datasets using **WhereScape**.
* Responsible for translating business and data requirements into logical data models in support of Enterprise data models, **ODS, OLAP, OLTP** and Operational data structures.
* Provided thought leadership for architecture and the design of **Big Data** Analytics solutions for customers, actively drive **Proof of Concept (POC)** and **Proof of Technology (POT)** evaluations and to implement a **Azure Data solution.**

**Environments**:Big data,Spark, Azure,TSQL,SQL,Metadata,Unix

**Client: Wawa- Allentown, PA Jan 2019 - Aug 2021**

**Role: Senior Data Engineer**

**Responsibilities:**

* Handled **Hadoop** **cluster** installations in **Windows** environment.
* Migrated on-premises environment in **GCP** **(Google Cloud Platform)**
* Migrated data warehouses to **Snowflake Data warehouse.**
* Involved in **Agile** Development process **(Scrum and Sprint planning).**
* Defined **virtual** **warehouse** sizing for **Snowflake** for different type of workloads.
* Integrated and automated data workloads to **Snowflake Warehouse.**
* Created tables in **snowflake DB,** loading and analyzing data using **Spark-Scala** scripts.
* Developed **ETL pipelines** in and out of data warehouse using combination of **Python** and **Snowflake’s Snow SQL.**
* Written **POCs** in **Python** to analyze the data quickly before applying **big data solutions** to process at a scale.
* Responsible for **data governance** rules and standards to maintain the consistency of the business element names in the different data layers.
* Build **Data Warehouse** in **Azure platform** using **Azure data bricks** and **data factory.**
* Developed **data pipeline** using **Sqoop** to ingest cargo data and customer histories into **HDFS** for **analysis**.
* Designed **ETL** using Internal/External tables and store in parquet format for efficiency.
* Involved in porting the existing on-premises **Hive** **code** **migration** to **GCP** **(Google Cloud Platform) BigQuery**.
* Involved in migration an **Oracle** **SQL** **ETL** to run on **Google cloud platform** using cloud **Dataproc** & **BigQuery**, cloud pub/sub for triggering the **Apache Airflow** jobs.
* Extracted data from **data lakes, EDW** to relational databases for analyzing and getting more meaningful insights using **SQL** **Queries** and **PySpark.**
* Used **AWS** services like **EC2** and **S3** for small data sets processing and storage, Experienced in Maintaining the **Hadoop** cluster on **AWS** **EMR**.
* Designed, developed and did maintenance of **data integration** programs in a **Hadoop** and **RDBMS** environment with both traditional and non-traditional source systems.
* Developed **MapReduce** programs to parse the raw data, populate staging tables and store the refined data in partitioned tables in the **EDW.**
* Wrote **Sqoop** **Scripts** for importing and exporting data from **RDBMS** to **HDFS**.
* Set up **Data Lake** in **Google cloud** using Google cloud storage, **BigQuery** and **Big** **Table.**
* Developed scripts in **BigQuery** and connecting it to reporting tools.
* Designed workflows using **Airflow** to automate the services developed for Change data capture.
* Carried out data transformation and cleansing using **SQL** **queries** and **PySpark**.
* Used **Kafka** and **Spark** streaming to ingest real time or near real time data in **HDFS**.
* Worked related to downloading **BigQuery** data into **Spark** **data** **frames** for advanced **ETL** capabilities.
* Built reports for monitoring data loads into **GCP** and drive reliability at the site level.
* Participated in daily stand-ups, bi-weekly scrums and PI panning.

**Environment:** Hadoop , GCP, BigQuery, Big Table, Spark , Sqoop , ETL, HDFS, Snowflake DW, Oracle Sql, MapReduce, Kafka and Agile process.

**Client: Prudential Financial- Newark, NJ Mar 2016 - Dec 2018**

**Role:Data Engineer**

**Responsibilities:**

* Worked as **Data** **Engineer** to review business requirement and compose source to target data mapping documents.
* Conducted technical orientation sessions using documentation and training materials.
* Gathered the business requirements from the Business Partners and Subject Matter Experts.
* Served as technical expert guiding choices to implement analytical and reporting solutions for client.
* Worked closely with the **business, other architecture team members** and global project teams to understand, document and design **data warehouse** processes and needs.
* Implemented Installation and configuration of multi-node cluster on Cloud using **Amazon Web Services (AWS)** on **EC2**.
* Developed reconciliation process to make sure **elastic search** index document count match to source records.
* Maintained **Tableau** functional reports based on user requirements.
* Created action filters, parameters, and calculated sets for preparing dashboards and worksheets in **Tableau.**
* Used **Agile (SCRUM)** methodologies for Software Development.
* Developed data pipelines to consume data from **Enterprise Data Lake (MapR Hadoop distribution - Hive tables/HDFS)** for analytics solution.
* Created **Hive** External tables to stage data and then move the data from Staging to main tables.
* Wrote **complex Hive queries** to extract data from heterogeneous sources **(Data Lake)** and persist the data into **HDFS**.
* Implemented the **Big Data** solution using **Hadoop, hive** and **Informatica** to pull/load the data into the **HDFS** system.
* Developed incremental and complete load **Python** processes to ingest data into **Elastic Search** from **Hive.**
* Pulled the data from data lake **(HDFS)** and massaging the data with various **RDD** transformations.
* Created **Oozie workflow** and Coordinator jobs to kick off the jobs on time for data availability.
* Developed Rest services to write data into **Elastic Search** index using **Python** Flask specifications
* Developed complete end to end **Big-data** processing in **Hadoop** eco system.
* Used **AWS Cloud** with Infrastructure Provisioning / Configuration.
* Used **Hive** to analyze the partitioned and **bucketed** data and compute various metrics for reporting on the dashboard.
* Created dashboards for analyzing **POS** data using **Tableau**.
* Developed **Tableau** **visualizations** and dashboards using **Tableau** **Desktop**.
* Involved in **PL/SQL** query optimization to reduce the overall run time of stored procedures.
* Used **Hive** to analyze the partitioned and **bucketed** data and compute various metrics for reporting on the dashboard.
* Continuously tuned **Hive UDF's** for faster queries by employing partitioning and **bucketing**.
* Implemented partitioning, dynamic partitions and buckets in **Hive**.
* Deployed **RMAN** to automate backup and maintaining scripts in recovery catalog.
* Worked on **QA** the data and adding **Data sources, snapshot, caching** to the **report.**

**Environment: AWS, Python, Agile, Hive, Oracle 12c, Tableau, HDFS, PL/SQL, Sqoop , Flume**

**Client: Netcrackers - India Sep 2013 - Oct 2015**

**Role: Software Developer**

**Responsibilities:**

* Responsible and involved in the start to end process of Hadoop cluster installation, configuration and monitoring.
* Responsible for building scalable distributed data solutions using Hadoop.
* Performed performance tuning and troubleshooting of Map Reduce jobs by analyzing and reviewing Hadoop log files.
* Developed several custom User defined functions in Hive & Pig using Java.
* Installed and configured Hadoop Map reduce, HDFS, Cassandra and developed multiple Map Reduce jobs in java for data cleaning and preprocessing.
* Work with hive complex data types and involved in Bucketing.
* Import millions of structured data from relational databases using Sqoop import to process using Spark and stored the data into HDFS in CSV format.
* Having experience on RDD architecture and implementing Spark operations on RDD and also optimizing transformations and actions in Spark.
* Developed UDF's using both Data Frames/Sql and RDD in Spark for data Aggregation queries and reverting back into OLTP through Sqoop.
* Developed Rest architecture-based web services to facilitate communication between client and servers.
* Used the Eclipse as IDE, configured and deployed the application into Web Logic application server using Maven build scripts to automate the build and deployment process.
* Implement Spark RDD transformations, actions to migrate Map reduce algorithms.
* Worked on NoSQL databases including HBase, MongoDB.
* Used Spark API over Horton works Hadoop YARN to perform analytics on data in Hive.

**Environment:** Hadoop, HDFS, Hive, Java, Sqoop, HBase, JIRA, GitHub

**Client: Revalsys Technologies - India Apr 2012 - Sep 2013**

**Role: Junior Developer**

**Responsibilities:**

* Effectively led client projects. These projects contained a heavy Python, SQL, Tableau and data modelling.
* Performed data merging, cleaning, and quality control procedures by programming data object rules into a database management system.
* Created detailed reports for management.
* Reported daily on returned survey data and thoroughly communicated survey progress statistics, data issues, and their resolution.
* Involved in Data analysis and quality check.
* Extracted data from source files, transformed and loaded to generate CSV data files with Python programming and SQL queries.
* Stored and retrieved data from data-warehouses.
* Created the source to target mapping spreadsheet detailing the source, target data structure and transformation rule around it.
* Wrote Python scripts to parse files and load the data in database, used Python to extract weekly information from the files, Developed Python scripts to clean the raw data.
* Worked extensively with Tableau Business Intelligence tool to develop various dashboards.
* Worked on datasets of various file types including HTML, Excel, PDF, Word and its conversions.
* Analysed data from company databases to drive optimization and improvement of product development, marketing techniques and business strategies
* Performed Database and ETL development per new requirements as well as actively involved in improving overall system performance by optimizing slow running/resource intensive queries.
* Developed data mapping documentation to establish relationships between source and target tables including transformation processes using SQL.
* Participated in data modelling discussion and provided inputs on both logical and physical data modelling.
* Reviewed the Performance Test results to ensure all the test results meet requirement needs.
* Created master Data workbook which represents the ETL requirements such as mapping rules, physical Data element structure and their description.

**Environment: Oracle 10g, UNIX Shell Scripts, MS Excel, MS Power Point, Python , SQL.**

**Client: W3Softech- India Jun 2011- Mar 2012**

**Role: Data Analyst**

**Responsibilities:**

* Preparation of unstructured data from multiple data files from customers database and adhoc
* reporting requests for customers as needed along with business review analysis of the program data.
* Communicated with the Source code provider in case of any discrepancies.
* Extracted, Transformed and Loaded (ETL) data to map it from disparate sources to the required target
* database.
* Worked on production support activities to monitor all Daily, Weekly, Monthly, Quarterly jobs and in
* Scheduler, fixing the failed workflows, communicating to the different teams to get the issue fixed
* based on the issue.
* Monitoring all Daily, Weekly, Monthly, Quarterly jobs and tracking the run statistics.
* Delved into data to discover discrepancies and patterns.
* Management and internal teams to implement and evaluate improvements.

**Environment: SQL Server, Oracle 10g, MS Excel.**