A blue hexagon with white text

Description automatically generatedA logo for a company

Description automatically generatedParveen Kumar

Sr. Data Engineer

Parveenkumare45@GMAIL.COM

+1-469-554-9624

<https://www.linkedin.com/in/p-mittal-72032525b/>

# Profile Summary

* **Azure certified professional with 10+** years of experience in **Data Engineering/ Data Architect, Big Data**/**ETL solution** development, **Data Visualization** & **Reporting**
* **Data Architect experience** with proven ability to bridge your business and technical teams, using a host of approaches, including AGILE and new databases, such as NoSQL.
* Extensive knowledge in working with **Azure cloud** platforms **(HDInsight, Data Lake, Data Bricks, Blob Storage, Data Factory, Azure Synapse, Azure Analytics, SQL, SQL DB, DWH, and Data Storage Explorer)**.
* Extensive experience in working with **NoSQL** databases and integration **with Dynamo DB, Cosmo DB, Mongo DB, Cassandra, and HBase.**
* Experienced in Designing, Developing, Documenting, and **Testing ETL** jobs and mappings in Server and Parallel jobs using Data Stage to populate tables in **Data Warehouse** and **Data marts**
* Hands-on experience in using **Hadoop ecosystem** components like **Hadoop, Hive, Pig, Sqoop, HBase, Cassandra, Spark, Spark Streaming, Spark SQL, Oozie, Zookeeper, Kafka, Flume, MapReduce framework, Yarn, Scala, and Hue**
* Experience **in ETL/ELT ingesting** data from various sources into Data Warehouses using business transformations and **star/snowflake data modeling**
* Experience in developing **CI/CD** (continuous integration and continuous deployment) and automation using **Git, ADO Board and Kubernetes**
* Expertise in **Data Migration, Data Profiling, Data Cleansing**, **Transformation, Integration, Data Import,** and **Data Export** using multiple **ETL tools** such as **Informatica Power Centre, SSIS**
* Extensively used **Python pytest, pyodbc, NumPy, MySQL dB, sqlite3, snowflake-python-connector,** and other packages
* Developed various **shell scripts, Scala scripts** and **Python scripts** to automate **Spark jobs** and **Hive jobs**
* Strong Knowledge of the Architecture of Spark distributed and parallel processing using **Spark SQL**, **Spark Data frames APIs**, and Spark execution framework
* Experience in **OLAP, OLTP**, Business Intelligence, and **Data Warehousing** with emphasis on ETL and Business Reporting needs
* Worked on Key Performance Indicators **KPIs**, design of star schema, and **snowflake schema** in Analysis Services **SSAS**
* Designed and implemented data warehouses and data marts using conformed **Facts & Dimensions, Slowly Changing Dimensions**, Surrogate Keys, Star Schema, Snowflake Schema, etc
* Experience in **Performance Tuning** and **Debugging** of existing ETL processes using **Informatica** tools
* Experience in creating Snowflake Views, **clustered and non-clustered Indexes**, Stored Procedures, Triggers,

**Window functions**, aggregate functions, and Table Value Functions using DDL, DML and PL/T-SQL

* Experience in working with **CSV, text, Excel, parquet, Jason** formats of data, and data from **RESTful API**
* Experience with **Requests**, **NumPy, SciPy, Matplotlib, HTTPLib2, Urllib2, Beautiful Soup, Data Frame**, and **pandas python libraries** during the development lifecycle such as planning , design ,Development , Testing , Deployment , Review,Launch .
* Experience in predictive **modeling**, A/B testing, **hypothesis testing**, ANOVA & **Time series analysis**

## (Exponential Smoothing, AR, MA, ARIMA

* Skilled in correlation and multivariate analysis, **causal analytics, and advanced Regression Modelling**

using statistical modeling concepts

* Experience in analyzing the data using MS Excel (**VBA, V - lookup, H – lookup, pivot tables, what-if analysis**, scenarios, link cells, charts) and MS Access
* Good understanding of **Machine learning algorithms** (Deep Neural Networks, Decision Trees, Linear & Logistic Regression, Random Forest, SVM, Bayesian statistics, XG Boost, K-Nearest Neighbors)
* Experience in data processing like collecting, aggregating, and moving data from various sources using

## AWS Glue creating Glue crawler jobs, Glue Catalogues, and Python scripts

* Worked on **AWS services** like **EC2, S3, EMR, RDS, Sage Maker, Redshift, DynamoDB, & Quick Sight** and other services of the AWS family
* Managed multiple tasks and worked under **tight deadlines and in a fast-paced environment**
* Team player, quick learner, organized and self-motivated

## TECHNICAL SKILLS

|  |  |
| --- | --- |
| **Languages** | **SQL (**Stored Procedures, User Defined Functions, **Window Functions, Triggers**, Views, Aggregated Functions, Table-valued Functions)  **Python (Pandas, NumPy, Scikit Learn, OOPS, Functional Programming)**  **Python Scripting, Shell scripting, Scala, SQL, Advanced Excel Scala, Scala Scripting, Shell Scripting** |
| **Version Control** | **GIT, GIT HUB, SVN, CVS, Azure Repositories** |
| **IDE & Tools, Design** | **Visual Studio, CI/CD, Jupiter Notebook, PyCharm**, **Databricks** |
| **Databases** | AWS S3, Azure Blob Storage, **MY SQL, MS SQL Server** |
| **Operating Systems** | **Linux, Unix, Mac OS-X, CentOS, Windows 10, Windows 8** |
| **Cloud Technologies** | **AWS**: AWS Sage Maker, **AWS Athena**, AWS Glue, AWS EMR, AWS RDS, Lambda, AWS CloudWatch, AWS DMS, **AWS Pipeline**, S3, EC2, **AWS Aurora**  **AZURE: Azure Data Factory, Azure Data Bricks**, Azure Data Lake and Azure Functions, **Azure Blob Storage**, Azure Runtimes, **ETL, ELT** |
| **Machine Learning Algorithms** | **Supervised** (Multiple Linear Regression, Logistic Regression, SVM, Stochastic Gradient Techniques, Decision Trees, Random Forest, Naïve Bays Algorithm)  **Unsupervised** (KNN, K-Means Clustering, Principal Component Analysis, t-SNE, LDA) |
| **ETL/ Reporting** | AWS Quick Sight, **Tableau, Power BI**, Matplotlib, Seaborn, **SSIS, SSRS**, **Informatica** |
| **Big Data Ecosystem** | **HDFS, MapReduce, Hive, Sqoop, Flume, Apache Spark, HBase,**  **Impala, Oozie, Spark, Zookeeper and Cloudera Manager, Kafka, Hadoop** |

**AWS Certified Solutions Architect – Associate**

<https://www.credly.com/badges/33efe5d7-eede-4ec9-8e11-59da47d21afc/linked_in_profile>

**Microsoft Certified: Azure Data Engineer Associate**

<https://learn.microsoft.com/en-us/users/kumarparveen-8432/credentials/d20c1fd297ebb6ae>

# Professional Experience

**Client Panoramic Health, AZ**  Aug 2021- Present **Designation: Senior Data Engineer**

**Responsibilities:**

* Developed 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 customer usage patterns.
* 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**.
* Built end-to-end **Azure Data factory pipelines** to process data from SQL server to **Azure Databricks** for time series forecasting across the different categories of products and segments
* Designed and implemented **metadata-driven data processing** frameworks, allowing for the automatic discovery and utilization of runtime parameters stored in metadata repositories.
* Leveraged metadata-driven approaches to seamlessly adjust **ETL behavior based** on changing requirements or environmental conditions, reducing manual intervention and increasing operational efficiency.
* Optimized **SQL stored procedures** and **SQL query performance** by execution plans optimization, **data partitioning,** and **data bucketing**
* Built the infrastructure required for optimal extraction, transformation, and loading of data from a wide variety of data sources using SQL and **Azure cloud** services
* Wrote and executed performance-tuned SQL stored procedures for **Data Analysis & Profiling,** and wrote complex SQL queries using joins, sub-queries, and correlated subqueries
* Built analytics solutions that utilize the data to provide actionable insights into customer acquisition, operational efficiency, **built KPIs** and other business performance metrics
* Compared reliability for different categories of products using **SQL, Power Pivot, Power BI,** and **Python**
* Built end-to-end **Azure Data factory** pipelines to process data from SQL server to Azure Databricks for time series forecasting across the different categories of products and segments

### Environment:

Azure Data Factory, Azure Utility **PL/SQL,** Azure SQL, Azure HDInsight, Databricks, MySQL, PostgreSQL, Snowflake, MongoDB, Cassandra, Teradata, Python, Spark, RDD, Spark, Azure Utility ,Spark SQL Airflow, Hive, Sqoop

**Client AT&T, TX**

**Designation: Data Engineer** Oct 2019 – Jul 2021

**Responsibilities:**

* Worked on Azure cloud platform services like **Azure Data Factory, Data Lake, Data Bricks**, **Blob Storage, Azure Analysis Services, Azure Synapse, SQL, SQL DB, DWH, and Data Storage Explorer**
* Ingested data to one or more Azure Services - (**Azure Data Lake, Azure Blob Storage, Azure SQL DW**) and processed the data using **Azure Databricks** and **Azure Analytics services**
* Designed and deployed data pipelines using **Azure Data Factory, Data Bricks, and Apache Airflow, Azure Utility**
* Created data pipeline for different events in Azure Blob storage into **Hive external tables** and used various

**Hive optimization** techniques like **partitioning, bucketing, and Map join**

* Worked on Azure Data Factory to integrate data of both on-prem (**MySQL, PostgreSQL, Cassandra**) and cloud (Blob Storage, Azure SQL DB) and applied transformations to load back to **Azure Synapse**
* Efficient in writing Python scripts to build **ETL pipelines** and **Directed Acyclic Graph (DAG) workflows using Apache Airflow, and Apache NiFi,Azure Utility**
* Designed custom-built input adapters using Spark, Hive, and Sqoop to ingest and analyze data **(Snowflake, MS SQL, MongoDB)** into HDFS
* Used **Azure DevOps (ADO Board)**and VSTS (Visual Studio Team Services) for **CI/CD, Active Directory** for authentication, and **Apache Ranger** for authorization
* Developed JSON Scripts for **deploying the Pipeline** in **Azure Data Factory (ADF)** that process the data using the SQL Activity

### Environment:

Azure Data Factory, Blob Storage, Synapse, Azure SQL, Azure HDInsight, Databricks, DataLakeGen2, Cosmos DB, AKS, MySQL, PostgreSQL, Snowflake, MongoDB, Azure Utility , Cassandra, Teradata, Python, Spark, RDD, PySpark, Airflow, DAG, Hive, Sqoop, HBase, AKS

**Client :** **American Express, Utah**

**Designation: Data Engineer** Jan 2018 – April 2019

**Responsibilities:**

* Involved in creating data ingestion pipelines for collecting data from various external sources like **FTP**

Servers and **AWS S3** buckets

* Designed and developed Security Framework to provide fine-grained access to objects in **AWS S3** using

## AWS Lambda, and DynamoDB

* Used **AWS EMR** to transform and move large amounts of data into and out of other AWS data stores and databases, such as **AWS S3 and Amazon DynamoDB**
* Worked on the code transfer of a quality monitoring program from **AWS EC2 to AWS Lambda**, as well as the creation of logical datasets to administrate quality monitoring on **Snowflake warehouses**
* Responsible for **monitoring and troubleshooting** the **AWS EMR Spark cluster**
* Developed ETL framework using **Spark and Hive** (including daily runs, error handling, and logging)
* Involved in building a real-time pipeline using **Kafka** and **Spark** streaming for delivering event messages to the downstream application team from an external rest-based application
* Worked on Automation and scheduling of **UNIX shell scripts** and batches using **Airflow bash operations**
* Worked on analyzing **spark cluster** and different big data analytic tools such as **HiveQL**
* Developed **Python and Scala scripts**, and UDFs using both **Data frames/SQL and RDDs** in Spark for **Data Aggregation**, queries, and writing data back into RDBMS
* Wrote shell scripts to monitor the health check of **spark clusters** and respond accordingly to any warning or failure conditions
* Used broadcast variables in **spark**, effective & efficient Joins, caching, and other capabilities for data processing
* Created **Partitioned Hive tables** (managed and external) and worked on them using **HiveQL**

### Environment:

AWS, Hadoop, Kafka, Airflow (MWAA), Unix, AWS EMR, Dynamo DB, Spark, AWS Lambda, HDFS, SQL, Spark, PySpark, Hive, HDFS, Sqoop, Kafka, HBase, Scala

## Client Verizon, VA

**Designation: ETL Developer** Nov 2015– Dec 2017

## Responsibilities:

* + Created **data dictionary, Data mapping for ETL and application support**, DFD, **ERD,** mapping documents, metadata, **DDL, and DML** as required
  + Created different types of reports including financial reports, **crosstab, conditional, drill-down, sub- reports** also **parameterized reports,** and **ad hoc reports** for existing databases
  + Design and created financial report templates, bar **graphs, and pie charts** based on the **financial data**. **Scheduled the monthly /weekly/daily reports** to run automatically
  + Encapsulated frequently executed SQL statements into **stored procedures, functions, indexes, and views**

to reduce the query execution times.

* + Designed, analyzed, and Implemented ETL processes, **ETL Testing**, ETL standards, and Naming conventions and wrote ETL flow documentation for Stage, ODS, and **Data Marts**
  + Used **SSIS** and **T-SQL** stored procedures to transfer data from **OLTP databases** to **the staging area** and finally transfer it into the **data warehouse**
  + Created SSIS packages to implement **error/failure handling** with **event handlers**, row redirects, and loggin.
  + Created SQL server configurations for SSIS packages and experienced in **creating jobs, alerts, and SQL mail agents** and **scheduling SSIS package.**
  + Used various **SSIS tasks** such as conditional split, derived column, and lookup which was used for data scrubbing, and data validation checks during Staging, and before loading the data into the **data warehouse.**
  + Responsible for creating reports based on the requirements using **SSRS** and defined the report layout and identified datasets for the report generation.

### Environment:

MS SQL Server, ETL, OLAP, OLTP, SQL, SSIS, SSRS, SSAS

**Client Syntel, New Delhi, India**

**Designation: Reporting Analyst**  Aug 2013– Oct 2015

**Responsibilities:**

* + Created reports in **Tableau for visualization** of the data sets created and tested Spark SQL connectors
  + Exporting a result set from **HIVE** to SQL using the Sqoop export tool for further processing.
  + Created **Partitioned Hive tables** (managed and external) and worked on them using **HiveQL**
  + Worked with **Tableau** and Integrated Hive, **Tableau Desktop** reports and published to **Tableau Server**
  + Create **Tableau interactive dashboards** and reports for teams within the organization
  + Converted Excel formulas into **VBA** code including **V - lookup, and H – lookup,** and developed macros for data automation
  + Actively involved in **User Acceptance Testing (UAT)** and Training of the end users. Participated in cross- functional teams to reengineer and improve business processes
  + **Validated and modeling data** with the purpose of understanding or making conclusions from the data for decision-making purposes
  + Maintain or updated thousands of contract **validity periods, updating, and posting bills receipts**, and invoices; ensuring all documents are accurate before issuing credits
  + Supported exchange of feedback and communication between management and associates
  + Logged reports, contracts, invoices, and sales documents into a **digital database** with Excel
  + Checked all layers of maps to promote accuracy, identifying and marking errors and making corrections
  + Compiled, organized, and corrected data with **advanced Excel** functions
  + Participated in meetings and Database Analysis with **Developers, Project Managers, and Quality Analysts**
  + to discuss business requirements, test planning, resource utilization, and defect tracking.
  + Developed **Data Flow Diagrams**, illustrating the flow of data from the legacy systems into the
  + Analyzing data using **statistical techniques** and providing reports.
  + Implemented **Tableau BI Dashboard** reporting solution for different groups in the organization