**DATA ENGINEER**

**Name: Shubham**

**Email:** [**shubham14shubham14@gmail.com**](mailto:shubham14shubham14@gmail.com)

**Contact: 1(469) 443-8004 // (609)-662-1275.**

**Phone:** [**https://www.linkedin.com/in/fnu-shubham-data-engineer/**](https://www.linkedin.com/in/fnu-shubham-data-engineer/)

**PROFESSIONAL SUMMARY:**

* IT professional with 9+ years of experience, specialized in Big Data ecosystem, Data Acquisition, Ingestion, Modelling, Storage Analysis, Integration, Data Processing, and Database Management.
* Experience in designing interactive dashboards, reports, performing ad-hoc analysis and visualizations using Tableau, Power BI, Arcadia, and Matplotlib.
* Experience in application development, implementation, deployment, and maintenance using Hadoop and Spark-based technologies like Cloudera, Hortonworks, Amazon EMR, and Azure HDInsight.
* Experienced in building data pipelines using Azure Data Factory, Azure Databricks, and loading data to Azure Data Lake, Azure SQL Database, Azure SQL Data Warehouse, and controlling database access.
* Extensive experience with Azure services like HDInsight, Stream Analytics, Active Directory, Blob Storage, Cosmos DB, and Storage Explorer. 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.
* Hands on experience in Architecting Legacy Data Migration projects such as Teradata to AWS Redshift migration and from on-premises to AWS Cloud.
* Use of NLP, Open NLP & Stanford NLP for Natural Language Processing, and sentiment analysis.
* Well versed with big data on AWS cloud services i.e., EC2, S3, Glue, Anthena, DynamoDB and RedShift
* Performed the migration of Hive and MapReduce Jobs from on - premises MapR to AWS cloud using EMR.
* Set up a Google Cloud Platform (GCP) Firewall rules to allow or deny traffic to and from the Virtual Machine instances based on specified configuration and used GCP cloud CDN (content delivery network) to deliver content from GCP cache locations drastically improving user experience and latency.
* Experience with Apache Spark ecosystem using Spark-Core, SQL, Data Frames, RDD's Spark MLlib.
* Strong experience in Business and Data Analysis, Data Profiling, Data Migration, Data Integration, Data governance and Metadata Management, Master Data Management and Configuration Management.
* Experience in implementing various Big Data Engineering, Cloud Data engineering, Data Warehouse, Data Mart, Data Visualization, Reporting, Data Quality, and Data virtualization Solution
* Experience in analysing data using Python, R, SQL, Microsoft Excel, Hive, PySpark, Spark SQL for Data Mining, Data Cleansing, Data Mining and Machine Learning.
* 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, and Azure Data Lake.
* Extensive Knowledge on developing Spark Streaming jobs by developing RDD’s (Resilient Distributed Datasets) using Scala, PySpark and Spark-Shell.
* Experience on Snowflake to manage, process, and analyse data efficiently in the cloud and it is scalable, cloud-based data warehousing and analytics solutions, enabling seamless data integration, transformation, and analysis.
* Good knowledge in understanding the security requirements and implementation using Azure Active Directory, Sentry, Ranger, and Kerberos for authentication and authorizing resources.
* Good knowledge of Hadoop cluster architecture and its key concepts - Distributed file systems, Parallel processing, High availability, fault tolerance, and Scalability.
* Complete knowledge of Hadoop architecture and Daemons of Hadoop clusters, which include Name node, Data node, Resource manager, Node Manager, and Job history server.
* Expertise in developing Spark applications for interactive analysis, batch processing and stream processing, using programming languages like PySpark, Scala.
* Advanced knowledge in Hadoop based Data Warehouse (HIVE) and database connectivity (SQOOP).
* Ample experience using Sqoop to ingest data from RDBMS - Oracle, MS SQL Server, Teradata, PostgreSQL, and MySQL.
* Experience in working with various streaming ingest services with Batch and Real-time processing using Spark Streaming, Kafka, Confluent, Storm, Flume, and Sqoop.
* Proficient in using Spark API for streaming real-time data, staging, cleaning, applying transformations, and preparing data for machine learning needs.

**TECHNICAL SKILLS**

|  |  |
| --- | --- |
| **Big Data Ecosystem** | HDFS, Yarn, MapReduce, Spark, Snowflake, Kafka, Kafka Connect, Hive, Airflow, Stream Sets, Sqoop, HBase, Flume, Pig, Ambari, Oozie, Zookeeper, NiFi, Sentry |
| **Hadoop Distributions** | Apache Hadoop 2.x/1.x, Cloudera CDP, Hortonworks HDP |
| **Cloud Environment** | Amazon Web Services (AWS), Microsoft Azure, GCP |
| **Databases** | MySQL, Oracle, Teradata, MS SQL SERVER, PostgreSQL, DB2 |
| **NoSQL Database** | DynamoDB, HBase |
| **AWS** | EC2, EMR, S3, Redshift, EMR, Lambda, Kinesis Glue, Data Pipeline |
| **Microsoft Azure** | Databricks, Data Lake, Blob Storage, Azure Data Factory, SQL Database, SQL Data Warehouse, Cosmos DB, Azure Active Directory |
| **Operating systems** | Linux, Unix, Windows 10, Windows 8, Windows 7, Windows Server 2008/2003, Mac OS |
| **Software’s/Tools** | Microsoft Excel, Stat graphics, Eclipse, Shell Scripting, ArcGIS, Linux, Jupiter Notebook, PyCharm, Vi / Vim, Sublime Text, Visual Studio, Postman |
| **Reporting Tools/ETL Tools** | Informatica, Talend, SSIS, SSRS, SSAS, ER Studio, Tableau, Power BI, Arcadia, Data stage, Pentaho |
| **Programming Languages** | Python (Pandas, SciPy, NumPy, Scikit-Learn, Stats Models, Matplotlib, Plotly, Seaborn, Keras, TensorFlow, PyTorch), PySpark, T-SQL/SQL, PL/SQL, HiveQL, Scala, UNIX Shell Scripting |
| **Version Control** | Git, SVN, Bitbucket |
| **Development Tools** | Eclipse, NetBeans, IntelliJ, Hue, Microsoft Office |

**Professional Experience**

**Client: HotelKey Inc, Carrollton, TX**  **Jan 2023 to Present**

**Data Engineer**

**Responsibilities:**

* Responsible for the execution of big data analytics, predictive analytics, and machine learning initiatives.
* Analysis, design and build Modern data solutions using Azure Cloud services to support visualization of data.
* Understand current Production state of application and determine the impact of new implementation on existing business processes.
* Extract Transform and Load data from the different Sources Systems to **Azure Data Storage services** using a combination of **Azure Data Factory, Spark SQL** and **Python** and **Azure Data Lake Analytics**.
* Data Ingestion to one or more Azure Services - (**Azure Data Lake, Azure Storage, Azure SQL, Azure DW**) and processing the data in In Azure Databricks.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to **Extract, Transform, and load data** from different sources like **Azure SQL**, **Blob storage**, **Azure SQL Data warehouse**, write-back tool and backwards.
* Responsible for **Azure Data Factory** job monitoring and troubleshooting the failures and providing the resolution for the ADF jobs failures.
* Developed Spark applications using **PySpark** and **Spark-SQL** for **data extraction, transformation, and aggregation** from multiple file formats for analysing & transforming the data to uncover insights into the customer usage patterns.
* Snowflake to ingest data from various sources such as databases, data lakes, streaming platforms, and third-party applications.
* Worked on Parquet file format and other kind of different file types. Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.
* Worked on a migration project to migrate data from different sources (**Teradata, Hadoop, DB2**) to **Google Cloud Platform (GCP)** using **UDP framework** and transforming the data using **Spark Scala scripts**.
* Worked on creating data ingestion processes to maintain Global Data Lake on the **GCP** cloud and Big Query.
* Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
* Worked on Production bugs especially involved in Azure Databricks Notebooks bugs and provided the new PySpark and Spark SQL logics to eliminate the bugs.
* Experience in Different kind of Data platforms (CUBE, DATAMARTS).
* Developed Notebooks and ETL Pipeline in Azure Data Factory (ADF) that process the data according to the job trigger.
* Hands-on experience on developing SQL Scripts for automation purpose.
* Created Build and Release for multiple projects (modules) in production environment using Visual Studio Team Services (VSTS).
* Worked on Google Cloud Platform (GCP) services like computer engine, cloud load balancing, cloud storage, cloud SQL, stack driver monitoring and cloud deployment manager.
* Strong understanding of Data Modeling in data warehouse environments such as star schema and snowflake schema.
* Experience in migrating the Legacy application into GCP platform and managing the GCP services such as Compute engine, cloud storage, Big Query, VPC, Stack Driver, Load Balancing and IAM.
* Worked on Spark Architecture including Spark Core, Spark SQL, Data Frames, Spark Streaming, Driver Node, Worker Node, Stages, Executors and Tasks.
* Involved in Database Design and development with Business Intelligence using SQL Server 2014/2016, Integration Services (SSIS), DTS Packages, SQL Server Analysis Services (SSAS), DAX, OLAP Cubes, Star Schema and Snowflake Schema.
* Extensively utilized SSIS packages to create complete ETL process and load data into database which was to be used by Reporting Services.
* Identified the dimension, fact tables and designed the data warehouse using star schema. Developed Multi-Dimensional Objects (Cubes, Dimensions) using SQL Server Analysis Services (SSAS).
* Extensively used Azure DevOps for code check-in and checkouts for version control.

**Environment:** Hadoop, Azure cloud Services (Azure Data Factory, Azure Data Bricks, Azure Data Lake), MS visual studio, GitHub, PySpark, Scala, Snowflake GCP, SQL Server, SQL, MS Power BI.

**Client: Tyroo Inc, India**  **Jan 2019 to Dec 2021**

**Data Engineer**

**Responsibilities:**

* Studied in-house requirements for the Data warehouse to be developed, conducted one-on-one sessions with business users to gather data warehouse requirements.
* Analysed database requirements in detail with the project stakeholders by conducting Joint Requirements.
* Development sessions developed a Conceptual model using Erwin based on requirements analysis.
* Developed normalized Logical and Physical database models to design OLTP system for insurance applications.
* Created dimensional model for the reporting system by identifying required dimensions and facts using Erwin r7.1.
* Used forward engineering to create a Physical Data Model with DDL that best suits the requirements from the Logical Data Model
* Giving on-call support to monitor and fixing the failures occurring in production and aware of various kinds of issues that comes up in both Ab initio and UNIX, Autosys Jobs.
* Designed and customized data models for Data warehouse supporting data from multiple sources on real time and Created mapping documents to outline data flow from sources to targets.
* Involved in building the ETL architecture and Source to Target mapping to load data into Data warehouse.
* Involved in Dimensional modelling (Star Schema) of the Data warehouse and used Erwin to design the business process, dimensions and measured facts.
* To execute SQL queries, user-defined functions (UDFs), stored procedures, and other computational tasks directly within Snowflake.
* Extracted the data from the flat files and other RDBMS databases into staging area and populated onto Data warehouse.
* Analyse, design and build Modern data solutions using Azure PaaS service to support visualization of data.
* Understand current Production state of application and determine the impact of new implementation on existing business processes.
* 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. Data Ingestion to one or more Azure Services - (Azure Data Lake, Azure Storage, Azure SQL, Azure DW) and processing the data in In Azure Databricks.
* Created Pipelines in ADF using Linked Services/Datasets/Pipeline/ to Extract, Transform, and load data from different sources like Azure SQL, Blob storage, Azure SQL Data warehouse, write-back tool and backwards.
* Installed and configured Apache airflow for workflow management and created workflows in python.
* Developed Spark applications using PySpark and Spark-SQL for data extraction, transformation, and aggregation from multiple file formats for analysing & transforming the data to uncover insights into the customer usage patterns.
* Responsible for estimating the cluster size, monitoring, and troubleshooting of the Spark data bricks cluster.
* Experienced in performance tuning of Spark Applications for setting right Batch Interval time, correct level of Parallelism and memory tuning.
* To meet specific business requirements wrote UDF’s in Scala and PySpark.
* Developed JSON Scripts for deploying the Pipeline in Azure Data Factory (ADF) that process the data using the SQL Activity.
* Hands-on experience on developing SQL Scripts for automation purpose.
* Created Build and Release for multiple projects (modules) in production environment using Visual Studio Team Services (VSTS).
* Well versed in SQL Server and T- SQL (DDL and DML) in constructing Tables, Normalization/ De normalization Techniques on database Tables.
* Involved in Creating and Updating Clustered and Non-Clustered Indexes to keep up the SQL Server Performance.
* Worked in creating and managing fragmentation of Indexes to achieve better query performance. Experience in Performance Tuning and Query Optimization.

**Environment**: Python, PySpark Data Virtualization, Data Warehouse, Snowflake, AWS, Hive, HBase, Impala, Airflow, Azure, ADF, Azure Data Lake (ADL), SQL Server, Sqoop, GCP, MapReduce, NOSQL, UNIX, HDFS, Oozie, SSIS.

**Client: Apeejay Stya Group Inc, India**  **June 2015 to Jan 2019**

**Data Engineer**

**Responsibilities:**

* Works closely with project manager to develop work plan for Data Warehouse projects and keep the manager aware of any issues.
* Supporting the development group by providing API's and performing middle-tier application Development.
* Providing analytical network support to improve quality and standard work results.
* Providing accurate estimates for project development and implementation. Work with management to meet those estimates.
* Develop framework, metrics and reporting to ensure progress can be measured, evaluated and continually improved.
* Perform in-depth analysis of research information for the purpose of identifying opportunities, developing proposals and recommendations for use by management.
* Support the development of performance dashboards that encompass key metrics to be reviewed with senior leadership and sales management.
* Well versed in designing, building, and implementing cloud systems.
* Hands on 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.
* Deployed application to GCP using Spinnaker (rpm based) and worked on designing star schema in Big Query.
* Deployed application to GCP using Spinnaker (rpm based) launched multi-node Kubernetes cluster in Google Kubernetes Engine (GKE) and migrated the dockerized application from AWS to GCP.

**Environment:** Cassandra, HDFS, MongoDB, Zookeeper, Oozie, Pig, Google Cloud Platform (scala), Kubernetes, GitHub, Jenkins, Big Query, Docker, JIRA, Unix/Linux CentOS 7, Nexus V3, Bash Shell Script, Python, Node.js, Apache Tomcat, MongoDB, SQL.